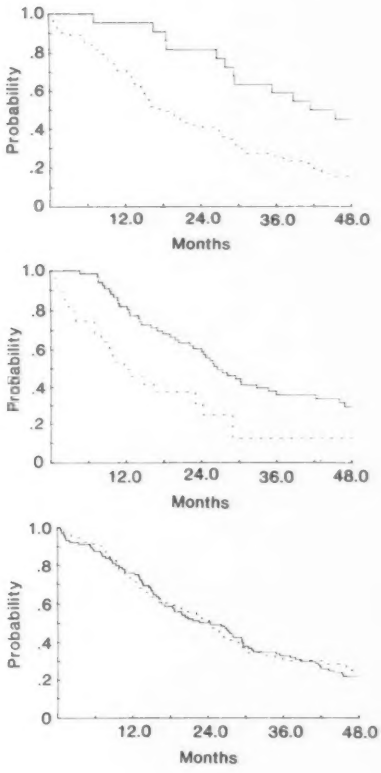
CANCER
TREATMENT REPORTS

Symposium on Methodology and Quality Assurance in Cancer Clinical Trials

October 1985
Volume 69
Number 10
Pages 1039–1233



U.S.Department of Health
and Human Services
Public Health Service
National Institutes of Health

CANCER
TREATMENT REPORTS

National Cancer Institute, Vincent T. DeVita, Jr, ***Director***

International Cancer Information Center
Susan Molloy Hubbard, ***Director***

Publications Branch
Jean Griffin Baum, ***Chief***

Robert E. Wittes, ***Editor-in-Chief***
Dee Kittner LeRoy, ***Managing Editor***

Editorial Board

Joseph Aisner
Lewis Aronow
Gregory Curt
Richard L. Cysyk
William D. DeWys
Ross C. Donehower
Michael Friedman
Timothy J. Kinsella
Robert F. Ozols
Carl Pinsky
David G. Poplack
Richard Simon

Advisory Board

Charles M. Balch
Robert C. Bast, Jr
Gianni Bonadonna
Edwin C. Cadman
Giulio J. D'Angio
I. David Goldman
Heine H. Hansen
Simon Kramer
Susan E. Krown
John E. Niederhuber
T. L. Phillips
H. M. Pinedo
Carol S. Portlock
Joseph V. Simone
Samuel A. Wells, Jr

Editorial Staff

Editors:
Elaine Price Beck
Maralyn Farber Berlin
Mary Olvey Cowgill

Symposium Editor:
Anne McCarthy
Editorial Assistant:
Julia Redmond

Manuscript Office:
Sally Bernstein
Kathleen Sager
Olga Sisson

Editorial Policy

The Editorial Board of *Cancer Treatment Reports* will consider manuscripts on original observations in laboratory and clinical research from all scientists. A manuscript is considered for publication with the understanding that it has not been published previously and is submitted exclusively to *Cancer Treatment Reports*. Opinions expressed by the authors are not necessarily those of the publisher or its editors. Scientific contributions in any of the following areas of investigation are solicited: surgery, radiotherapy, chemotherapy, immunotherapy, supportive care, pharmacology, cell biology and kinetics, mechanisms of drug action, and medicinal chemistry.

Submission of Manuscripts

Manuscripts and all inquiries concerning the status of manuscripts should be addressed to: Managing Editor, Cancer Treatment Reports, R. A. Bloch International Cancer Information Center, Bldg 82, Rm 231, National Cancer Institute, Bethesda, MD 20892 (301-496-6975). All material submitted for consideration in the journal will be subject to review when appropriate by at least two outside reviewers and one member of the Editorial Board. The decision to publish a manuscript is solely the responsibility of the Editorial Board. Prospective authors should consult the "Instructions to Authors," which appears on the second page of every issue.

Subscriptions

Cancer Treatment Reports is for sale by the Superintendent of Documents, US Government Printing Office, Washington, DC 20402.

INSTRUCTIONS TO AUTHORS

Title Page
Includes complete title of the manuscript and the full name, affiliation, and address of each author. Indicate full name and address of author to whom correspondence and/or reprint requests are to be sent.

Summary
Begin the article with a short summary.

Text
Submit an original and 3 copies of the manuscript, typed double-spaced. Number each page consecutively. Dot-matrix or italicized texts will not be accepted. Use accepted spelling, abbreviation, and hyphenation found in *Webster's New International Dictionary*. Footnotes should be typed at the bottom of the page on which they appear.

Tables
Submit each table on a separate page, typed double-spaced. Please use only standard abbreviations; otherwise use full words.

References
Submit references typed double-spaced. Follow the standard form of *Index Medicus*: author, title, name of journal, volume number, page or inclusive pages, and year. In book references give author, chapter title, book title, city and name of publisher, date of publication, volume number, and pages cited. References in text should begin with number 1 and continue consecutively. References to papers published in meeting proceedings should include author; title of paper; title of meeting; city and name of publisher; date of publication; volume number; and pages. Papers presented at meetings but not published should be cited in a footnote rather than in the reference section.

Illustrations
Photographs and graphs should be clear, glossy prints. If original art work is submitted, it should be done in black ink for good reproduction. Please send 2 copies of each print. Photocopies reproduce poorly. Color photographs or transparencies are acceptable.

Nomenclature
Names of compounds should conform to American usage. See *USAN and the USP Dictionary of Drug Names* or *The Merck Index* for proper nomenclature.

Brief Reports
Manuscripts published in the Brief Reports section should be no longer than 5 double-spaced pages including references, tables, and a short summary; figures are usually omitted.

Letters to the Editor
Letters to the Editor should be no longer than 3 double-spaced text pages and should be submitted with no more than 1 table and 5 references; figures are usually omitted.

Clinical Trials Summaries
Manuscripts published in this section should be no longer than 300 words and should be submitted with no more than 1 table and 5 references, and no figures. Tables should include the following: *patient characteristics and prognostic factos* (age/sex, performance status, stage of disease, pathologic types of disease, prior therapy), *drug regimen*, or *standardized toxicity grades* to demonstrate that an appropriate, biologically active dose was administered.

Reprints
Supplied free of charge, 100 reprints (20 reprints for Letters to the Editor) will be sent to the senior or specified author about 3 weeks after publication of the article.

Announcements
Announcements of meetings are accepted for publication free of charge. Mail to: Editorial Office, Cancer Treatment Reports, R. A. Bloch International Cancer Information Center, Bldg 82, Rm 231, National Institutes of Health, Bethesda, MD 20205.

The Secretary of Health and Human Services has determined that the publication of this periodical is necessary in the transaction of the public business required by law of this Department. Use of funds for printing this periodical has been approved by the Director of the Office of Management and Budget through June 30, 1987.

METHODOLOGIC GUIDELINES FOR REPORTS OF CLINICAL TRIALS*

1. Authors should discuss briefly the quality control methods used to ensure that the data are complete and accurate. A reliable procedure should be cited for ensuring that all patients entered on study are actually reported upon. If no such procedures are in place, their absence should be noted. Any procedures employed to ensure that assessment of major endpoints is reliable should be mentioned (eg, second-party review of responses) or their absence noted.

2. All patients registered on study should be accounted for. The report should specify for each treatment the number of patients who were not eligible, who died or withdrew before treatment began. The distribution of follow-up times should be described for each treatment, and the number of patients lost to follow-up should be given.

3. The study should not have an inevaluability rate for major endpoints of greater than 15%. Not more than 15% of eligible patients should be lost to follow-up or considered inevaluable for response due to early death, protocol violation, missing information, etc.

4. In randomized studies, the report should include a comparison of survival and/or other major endpoints for all eligible patients as randomized, that is, with no exclusions other than those not meeting eligibility criteria.

5. The sample size should be sufficient to either establish or conclusively rule out the existence of effects of clinically meaningful magnitude. For "negative" results in therapeutic comparisons, the adequacy of sample size should be demonstrated by either presenting confidence limits for true treatment differences or calculating statistical power for detecting differences. For uncontrolled phase II studies, a procedure should be in place to prevent the accrual of an inappropriately large number of patients, when the study has shown the agent to be inactive.

6. Authors should state whether there was an initial target sample size and, if so, what it was. They should specify how frequently interim analyses were performed and how the decisions to stop accrual and report results were arrived at.

7. All claims of therapeutic efficacy should be based upon explicit comparison with a specific control group, except in special circumstances where each patient is his own control. If nonrandomized controls are used, the characteristics of the patients should be presented in detail and compared to those of the experimental group. Potential sources of bias should be adequately discussed. Comparison of survival between responders and nonresponders does not establish efficacy and should not generally be included. Reports of phase II trials which draw conclusions about antitumor activity but not therapeutic efficacy generally do not require a control group.

8. The patients studied should be adequately described. Applicability of conclusions to other patients should be carefully dealt with. Claims of subset-specific treatment differences must be carefully documented statistically as more than the random results of multiple-subset analyses.

9. The methods of statistical analysis should be described in detail sufficient that a knowledgeable reader could reproduce the analysis if the data were available.

**From* Methodologic Guidelines for Reports of Clinical Trials (Cancer Treat Rep 69:1–3, 1985).

Contents

Cancer Treatment Reports
Volume 69, Number 10
October 1985

Symposium on Methodology and Quality Assurance in Cancer Clinical Trials

1039 Introduction. **S. S. Ellenberg, R. Simon, and R. E. Wittes.** National Cancer Institute, Bethesda, Maryland

SESSION I: Biometric Issues (Dr. Richard Simon, chairperson)

1041 Role of Mathematical Modeling in Protocol Formulation in Cancer Chemotherapy. **A. J. Coldman and J. H. Goldie.** Cancer Control Agency of British Columbia, Vancouver, Canada

1047 DISCUSSION

1049 Response Surface Methodology and the Design of Clinical Trials for the Evaluation of Cancer Chemotherapy. **W. H. Carter, Jr.** Medical College of Virginia, Richmond

1053 DISCUSSION

1055 Factorial Designs for Randomized Clinical Trials. **D. P. Byar and S. Piantadosi.** National Cancer Institute, Bethesda, Maryland

1063 DISCUSSION

1065 Tree-Structured Survival Analysis. **L. Gordon and R. A. Olshen.** University of Southern California, Los Angeles

1069 DISCUSSION

1071 Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions. **F. E. Harrell, Jr, K. L. Lee, D. B. Matchar, and T. A. Reichert.** Duke University Medical Center, Durham, North Carolina

1079 PANEL DISCUSSION

SESSION II: General Issues (Dr. Susan Ellenberg, chairperson)

1087 Size of Phase III Cancer Clinical Trials. **R. Simon.** National Cancer Institute, Bethesda, Maryland

1093 DISCUSSION

1095 The Doctor's Dilemma: Physician Participation in Randomized Clinical Trials. **K. M. Taylor.** University of Toronto and Women's College Hospital, Canada

1099 DISCUSSION

1101 Policies for Interim Analysis and Interim Reporting of Results. **J. R. O'Fallon.** Mayo Comprehensive Cancer Center, Rochester, Minnesota

1105 DISCUSSION

1107 Eligibility Exclusions, Losses to Follow-up, Removal of Randomized Patients, and Uncounted Events in Cancer Clinical Trials. **M. H. Gail.** National Cancer Institute, Bethesda, Maryland

1113 DISCUSSION

1115 Measuring Quality of Life: Risks and Benefits. **H. Schipper and M. Levitt.** Manitoba Cancer Treatment and Research Foundation, Winnipeg, Canada

1125 DISCUSSION

1127 Influence of Measurement Error on Response Rates. **D. Warr, S. McKinney, and I. Tannock.** Princess Margaret Hospital and Ontario Cancer Institute, Toronto, Canada

1131 DISCUSSION

1133 PANEL DISCUSSION

SESSION III: Topics in Drug Development (Dr. Michael Friedman, chairperson)

1139 Analysis and Interpretation of the Comparison of Survival by Treatment Outcome Variables in Cancer Clinical Trials. **J. R. Anderson, K. C. Cain, R. D. Gelber, and R. S. Gelman.** Dana-Farber Cancer Institute, Boston, Massachusetts

1145 DISCUSSION

1147 An Efficient Design for Phase III Studies of Combination Chemotherapies. **S. S. Ellenberg and M. A. Eisenberger.** National Cancer Institute, Bethesda, Maryland

1153 DISCUSSION

1155 Food and Drug Administration Requirements for Approval of New Anticancer Drugs. **J. R. Johnson and R. Temple.** Food and Drug Administration, Rockville, Maryland

1159 DISCUSSION

SESSION IV: Biological Response Modifier Trials (Dr. Ronald Herberman, chairperson)

1161 Design of Clinical Trials With Biological Response Modifiers. **R. B. Herberman.** National Cancer Institute-Frederick Cancer Research Facility, Maryland

1165 Assessment of Biological Responses: What To Measure and When. **A. E. Maluish and R. B. Herberman.** National Cancer Institute-Frederick Cancer Research Facility, Maryland

1171 Applicability of Phase I Trial Results in the Design of Phase II and III Biological Response Modifier Trials. **C. M. Pinsky.** Memorial Sloan-Kettering Cancer Center and Cornell University Medical College, New York, New York

1175 GENERAL DISCUSSION

SESSION V: Quality Assurance (Dr. Luz Hammershaimb, chairperson)

1177 Site Visit Monitoring Program of the Clinical Cooperative Groups: Results of the First 3 Years. **J. K. Mauer, D. F. Hoth, D. K. Macfarlane, L. D. Hammershaimb, and R. E. Wittes.** National Cancer Institute, Bethesda, Maryland

1189 Quality Assurance in a Cooperative Group. **J. M. Wolter.** Rush-Presbyterian-St. Luke's Medical Center, Chicago, Illinois

1195 Clinical Quality Assurance in the Pharmaceutical Industry. **J. J. Donahue.** Smith Kline and French Laboratories, Philadelphia, Pennsylvania

1199 Quality Assurance of Radiotherapy in Clinical Trials. **A. S. Glicksman, L. E. Reinstein, and F. Laurie.** Rhode Island Hospital and Brown University, Providence

1207 Pathology Quality Control in the Cooperative Clinical Cancer Trial Programs. **R. L. Kempson.** Stanford University School of Medicine, California

1211 Prognostic Implications of Blast Cell Morphology in Childhood Acute Lymphoblastic Leukemia: A Report From the Childrens Cancer Study Group. **D. R. Miller, M. Krailo, W. A. Bleyer, J. N. Lukens, S. E. Siegel, P. R. Coccia, J. Weiner, and D. Hammond.** Childrens Cancer Study Group

1223 PANEL DISCUSSION

1231 ANNOUNCEMENTS

Symposium on Methodology and Quality Assurance in Cancer Clinical Trials

Sponsored by:
Cancer Therapy Evaluation Program
National Cancer Institute

Hyatt Regency
Washington, DC
October 24–26, 1984

INTRODUCTION

Progress in cancer treatment depends in an obvious way upon advances in biomedical science. If conceptual breakthroughs or sheer serendipity results in a new treatment that is enormously better than standard therapy, the advantages of the new treatment should shine through even a defective clinical trial. Unfortunately, such advances have been uncommon in the developmental therapeutics of cancer. Much more commonly we have had to settle for small or moderate improvements. When investigators have been clever enough to exploit a collection of individually modest advances, striking progress has sometimes emerged.

The ascertainment of relatively small differences in therapy, therefore, is of real importance, and under such circumstances attention to the details of clinical trials methodology becomes critical. Over the past few years the clinical research community has become increasingly sensitive to the importance of methodological issues. In September 1983, the National Cancer Institute and the National Heart, Lung, and Blood Institute jointly sponsored a Workshop on Evaluation of Therapy.[1] The presentations and discussion at this meeting dealt with a variety of general issues relating to trials methodology, including virtues and problems of randomized trials, the possible uses of observational data bases, and issues concerning the size, complexity, and efficiency of trials. At the conclusion of this stimulating meeting, many of us felt the need for another workshop devoted to some of the problems peculiar to cancer trials. The present volume of *Cancer Treatment Reports* contains the proceedings of this second meeting, which took place in Washington, DC, on October 24-26, 1984.

The subjects of this symposium reflect many concerns of thoughtful clinical oncologists, whether they are engaged in trials themselves or are attempting to glean from the literature information about the best treatment for individual patients. Papers in the various sessions deal with the impact of quantitative methods on study design, problems in the execution and interpretation of clinical trials, certain dilemmas in clinical drug development, and the unusual methodologic demands of trials with biologic response modifiers. We decided to reproduce the discussion sessions and panels which followed the formal presentations, since the informal interchanges were frequently as illuminating as the papers themselves.

The topics covered in this symposium by no means include all the issues in methodology that are relevant for the clinical oncologist. We plan to sponsor regular workshops on this general subject. We hope that these meetings will clarify some of the complex issues in clinical trials methodology and will lead to a wider realization that in the evaluation of therapy the question "What do we know?" cannot be divorced from the question "How do we know what we know?"

Susan S. Ellenberg
Richard Simon
Robert E. Wittes

[1]Friedman L, Simon R, Verter J, et al, eds. Proceedings of the Workshop on Evaluation of Therapy, September, 1983. Statistics and Medicine 3:305-486, 1984.

Role of Mathematical Modeling in Protocol Formulation in Cancer Chemotherapy

A. J. Coldman* and J. H. Goldie[1]

The origin of drug resistance is discussed and the spontaneous mutation model is presented. The general implications of this model for the use of chemotherapy are considered and are found to agree with the broad range of experience from treating human malignancy. The large amount of information required to use this model to create detailed protocols does not permit its utilization at present, but it is hoped that such information will become available in the future. [Cancer Treat Rep 69:1041–1045, 1985]

Resistance is a term generally used in radiobiology and chemotherapy to describe the inability of treatment to cure a person's disease. With the increased availability of a large variety of cytotoxic agents, it has become recognized as an essential obstacle to be overcome in the therapy of malignancy. Clinically this is most apparent when therapy with an agent is effective at first in reducing tumor volume, but subsequently, perhaps during the course of therapy, the tumor regrows, and continued therapy with this agent is no longer useful.

Extensive work in both animal tumor systems and humans has identified a large variety of mechanisms which can result in tumor resistance. One particular manifestation of resistance in which we have been particularly interested is that displayed by the individual tumor cell itself. In many situations, it has been recognized that some cells of the tumor are inherently resistant to a particular agent, while other cells in the tumor are sensitive to the agent. In vivo experimentation with several tumor lines by Skipper, Schabel (1,2), and others has shown that cells taken from tumors which are resistant maintain their resistance when transferred to other hosts. This suggests that some mechanisms of resistance are not properties of the host. Further work by these researchers on these transplanted resistant tumors has shown that in many cases, the kinetics of growth of the resistant tumor line are not substantially different from those of the parent sensitive line from which they were derived, and from this it would appear that chemoresistance can be due to mechanisms unrelated to tumor growth.

More recently, in vivo and in vitro research has shown that such phenomena are seen in a variety of tumor systems treated with a variety of chemotherapeutic agents. There is now overwhelming evidence that resistance can result from genetic changes occurring in the cell which are conserved at mitosis and lead to the development of cell lines which are resistant to chemotherapy (3). The present situation in cancer chemotherapy has strong parallels with the early experience in the chemotherapy of tuberculosis in the 1940s and 1950s. At that time, chemotherapies of increased efficacy were becoming available, but a substantial number of patients were not cured because of the overgrowth of resistant strains of the tuberculosis bacteria. In the 1930s and 1940s, there was much interest in the origins of resistant bacteria, and a number of experiments were undertaken to elucidate the mechanisms by which they came into being. In their classic experiments, Luria and Delbruck (4) were able to show that the origin of resistant types was compatible with a random mutational mechanism which was preserved at mitosis. The method they developed, the fluctuation test, has since been used in the analysis of resistance to cancer chemotherapy agents, and the results from such measurements are compatible with a similar origin for resistance to anticancer drugs (5,6).

Using this mechanism as a departure point, we have been interested in the mathematic modeling of the effect of these phenomena on the therapy of human malignancy. Following Luria and Delbruck's work, we have used the following assumptions to create a simplified model of tumor resistance.

[1]Biometry Section, Division of Epidemiology (*A. J. Coldman*), and Division of Medical Oncology (*J. H. Goldie*), Cancer Control Agency of British Columbia, Vancouver, Canada.

**Reprint requests to:* A. J. Coldman, MD, Biometry Section, Division of Epidemiology, Cancer Control Agency of British Columbia, 600 W 10th Ave, Vancouver, British Columbia, V5Z 4E6, Canada.

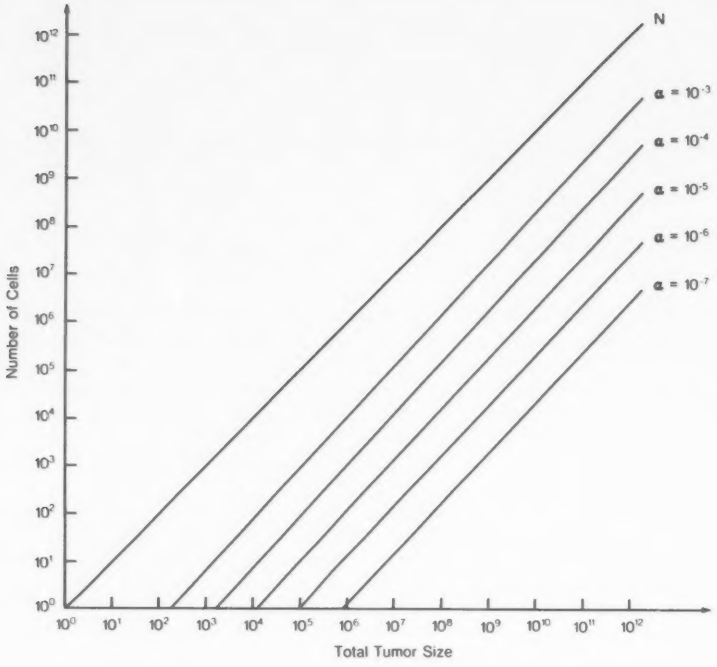


FIGURE 1.—Figure plots the expected No. of singly resistant cells as a function of the mutation rate and tumor size. The function is plotted for mutation rates ranging between 10^{-3} and 10^{-7}.

(*a*) All tumor cells are kinetically similar and have the ability for unlimited proliferation.

(*b*) Sensitive tumor cells have a fixed probability, a, per division of acquiring resistance to a particular drug at a given dose (a is frequently referred to as the mutation rate).

(*c*) Sensitive cells have a fixed probability of surviving the administration of an active chemotherapeutic agent (the log-kill law) (1), for cells resistant to an agent survive administration of that agent with probability one.

Before commenting further, it is important to reiterate that this model does not consider many other parameters which are important in the chemotherapy of human tumors. Normal tissue toxicity, which perhaps prevents effective chemotherapy more than any other factor, is not considered here. Cell cycle and other kinetic aspects are not considered and neither are pharmacologic sanctuaries of metastasis. Therefore, any deductions made which relate to the therapy of human tumors must be tempered, especially where they contradict considerations based on these other mechanisms for tumor insensitivity. Nevertheless, before considering possible, and perhaps necessary, refinements to the above model, it is interesting to examine deductions which can be made from this model and compare them with experience in experimental tumor systems and information gained from the therapy of human malignancies.

When considering the above model, it is a simple exercise to calculate the mean number of resistant cells as a function of the overall tumor size. This is plotted in figure 1 for a number of values of the mutation rate a. However, this quantity, although of interest in the calculation of the expected length of remission, is of comparatively little use in determining the overall long-term cure rate to be expected in the therapy of malignancy. A quantity of some interest is the probability that there are no resistant cells, since under this model this will translate into the likelihood that long-term cure can be achieved by the application of the chemotherapeutic agents. This function is given mathematically by $\exp(-a(N-1))$, and, as expected, is inversely related to the mutation rate and the overall tumor size. From figure 2, it may be seen that the shape of the function is fairly similar for different values of mutation rate, but is shifted in location.

It is possible to make a number of simple deductions

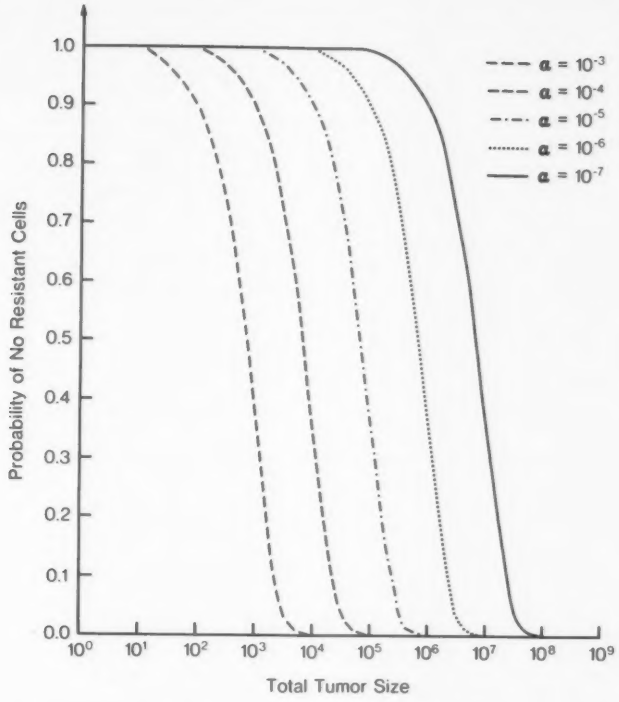


FIGURE 2.—Probability of no resistant cells is given as a function of the mutation rate and tumor size. The plots give values using mutation rates between 10^{-3} and 10^{-7}. For this range of mutation rates, the shape of each curve is quite similar, but their location varies with the mutation rate.

from this curve regarding the application of therapy.

(*a*) Other things being equal, earlier therapy, when the tumor is smaller, will have a greater likelihood of preventing the development of resistant cells.

(*b*) The probability that there are resistant cells present moves from high value to a low value over a 2-log increase in size whose location depends on the value of the mutation rate.

(*c*) For mutation rates in the range 10^{-3} to 10^{-7} (rates which have been determined in experimental systems), resistant cells will be present when disease is clinically detectable.

(*d*) Tumors will convert from a state of chemosensitivity to chemoresistance (one or more resistant cells present) in a period comparatively short compared to their total lifetime.

Although most of these results cannot be rigorously tested in human cancers, none of them appears to contradict the majority of experience with human malignancy. Furthermore, many of these results have been checked in in vivo experimental systems and appear to be compatible with much of the data collected there, notably on the L1210 leukemia, where a comparison of the predictions of this model and experimental results have been compiled by Skipper (7,8).

In human malignancy it has been known for some time that single-agent chemotherapy is unlikely to be effective except in a relatively small number of tumors. The combination of several active agents into single regimens has almost always been shown to have greater activity than any single member of a regimen alone. A possible method of incorporating multidrug regimens into the above framework would be to consider a single overall mutation rate to drug resistance α to the total regimen and proceeding as before. Evidence from some experimental tumors indicates that resistance to several drugs arises in a stepwise manner by the accumulation of single-drug resistance (6). From this observation, it is possible to mathematically calculate the overall mutation rates to resistance for various drug combinations (9). These calculations require a variety of assumptions, and, for complex drug combinations, involve the numeric specification of a large number of parameters. However, other lines of experi-

mental evidence indicate that resistance to single drugs may also involve resistance to a range of pharmacologically unrelated compounds (6). Experience in the therapy of human malignancy would seem to indicate that this lack of cross-sensitivity is also seen extensively in human tumors, although confounding with other factors is possible. For example, Skipper (10) has shown that in vivo tumors displaying resistance to CAF (cyclophosphamide, doxorubicin, and 5-FU) are only resistant to doxorubicin and not to the other two drugs. His data indicated that the doses of cyclophosphamide and 5-FU used in this combination were insufficient to retard the growth of doxorubicin-resistant cells, and that this was the eventual cause of treatment failure.

The definition and measurement of resistant cell lines imply necessarily that the frequency of cells resistant to several drugs must be no larger than the frequency of the cells resistant to the single drugs in that combination. This necessarily implies that drug combinations must be as effective against resistant sublines as any of the agents used singly in the same dose, as they are in the combination. This observation is thus in accord with the experience gained in the treatment of a large range of human malignancy.

Creating Protocols

The above considerations generally lead to two conclusions:

(*a*) Therapy should be given as early in the course of disease as possible.

(*b*) Combination chemotherapy should be as good as, and perhaps better than, single-agent therapy. This has led to the prediction that when two agents or two combinations of agents may not be given together, appropriate sequencing should prove at least as effective in controlling resistance of populations as the application of these regimens in any other fashion. When the two regimens are of equal effectiveness, then the correct scheduling is to sequentially alternate them (9). It is perhaps worth pointing out at this point that some currently used combination chemotherapy regimens could also be considered as alternating regimens when viewed appropriately. For example, consider MOPP (mechlorethamine, vincristine, procarbazine, and prednisone). In this case, procarbazine and prednisone are given on Days 1-14 and mechlorethamine and vincristine are given on Days 1 and 8, and therapy is repeated every 28 days. Viewed somewhat differently, we could consider this regimen as one where all drugs are given on Day 1 followed by a regimen of 6 days of administration of procarbazine and prednisone, whereupon the same regimen is repeated. Thus, protocols where two established therapeutic schedules are alternated can be viewed as the sequential administration of a regimen which includes both the previous ones. Therefore, it seems reasonable to view so-called alternating schedules as special cases of combination chemotherapy and expect that some may be successful and some may not, as has been the case with combination chemotherapy.

A particular protocol employing the principles previously discussed has been the development of chemotherapy regimens for the treatment of Hodgkin's disease. Previously, it has been shown that alternating MOPP and ABVD (doxorubicin, bleomycin, vinblastine, and dacarbazine) has led to greater success than using either of these two regimens alone (11). Clinical investigators at the Cancer Control Agency of British Columbia have developed a protocol called MOPP/ABV hybrid, which attempts to concentrate the two active regimens into a single regimen (table 1) (12). To date, this protocol has been reasonably successful, although results from a randomized trial are not yet available. The two obvious features of this protocol in comparison to MOPP and ABVD alone are the deletion of dacarbazine and the reduction in the length of administration of procarbazine.

However, this model does not solve, but merely emphasizes, the central problems: what drugs? what doses? It is a simple matter to state that use of the maximum variety of non-cross-resistant drugs should produce the greatest control of resistant subtypes. Host toxicity prevents unlimited combination of effective drugs; therefore, one is presented with a choice of possibly limiting dosage of certain drugs to increase the variety of agents which may be administered simultaneously. Although rather little data are available on

TABLE 1.—Regimen for MOPP/ABV (hybrid)

Drug (mg/m^2)	Days of therapy—														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15-28
Mechlorethamine (6)	X														
Vincristine (1.4) (maximum 2)	X														
Procarbazine (100)	X	X	X	X	X	X	X								
Prednisone (40)	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
Doxorubicin (35)								X							
Bleomycin (10)								X							
Vinblastine (6)								X							

this subject, the preponderance of evidence indicates that the doses of individual drugs should not be compromised in order to add other active agents.

In conclusion, it would seem that the most promising avenue for protocol formulation is in the search for drugs which are both active and non-cross-resistant. The possibility exists in the future that agents will be available which act specifically on cells resistant to chemotherapy, and these seem to hold the greatest benefit in the future for overcoming drug resistance. Mathematical modeling may be useful in suggesting general avenues for improvement in protocol design, but, as yet, the theory and the detailed data required are insufficient to permit detailed protocol specifications. It is hoped that in the future new techniques, such as human tumor cloning assays, will permit the specification of many of the parameters required so that mathematical models may be used.

REFERENCES

1. SKIPPER HE, SCHABEL FM, JR, and WILCOX WS. Experimental evaluation of potential anticancer agents. XIII. On the criteria and kinetics associated with curability of experimental leukemia. Cancer Chemother Rep 35:1-111, 1964.
2. SKIPPER HE. Reasons for success and failure in treatment of murine leukemias and the drugs now employed in treating human leukemias. *In* Cancer Chemotherapy. Ann Arbor, MI, University Microfilms International, 1978, vol 1, pp 1-66.
3. GOLDIE JH, and COLDMAN AJ. The genetic origin of drug resistance in neoplasms: implications for systemic therapy. Cancer Res 44:3643-3653, 1984.
4. LURIA SE, and DELBRUCK M. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28:491-511, 1943.
5. LAW LW. Origin of the resistance of leukaemic cells to folic acid antagonists. Nature 169:628-629, 1952.
6. LING V. Genetic basis of drug resistance in Mammalian cells. *In* Drug and Hormone Resistance in Neoplasia (Bruchovsky, N, and Goldie JH, eds). Boca Raton, C.R.C. Press, 1982, pp 1-19.
7. SKIPPER HE. Some thoughts regarding a recent publication by Goldie and Coldman. Southern Research Institute, Booklet No. 9, 1980.
8. SKIPPER HE. Additional thoughts on the model of Goldie and Coldman. Southern Research Institute, Booklet No. 10, 1980.
9. COLDMAN AJ, and GOLDIE JH. A model for the resistance of tumor cells to cancer chemotherapeutic agents. Math Biosci 65:291-307.
10. SKIPPER HE. Mammary adenocarcinoma 16/C: remission followed by relapse during continuing undiminished treatment with a three drug combination (CAF) delivered simultaneously. Southern Research Institute, Booklet No. 3, 1983.
11. SANTORO A, BONADONNA G, BONFONTE V, ET AL. Alternating drug combinations in the treatment of advanced Hodgkin's disease. N Engl J Med 306:770-775, 1982.
12. KLIMO P, and CONNORS JM. MOPP/ABV (HYBRID) in the treatment of advanced or recurrent Hodgkin's disease (initial experience). Proc ASCO 2:208, 1983.

DISCUSSION

Q: (Dr. Richard Simon, NCI) When you're trying to develop a protocol based upon your model, what estimates do you use for cross-resistance of the agents or the log kill of individual agents?

A: (Dr. Coldman) We've only really restricted modeling to the experimental tumors, so there we've been able to get fairly good data on log kills, on mutation rates, and, to a lesser extent, on cross-resistance. What we've attempted to do is to try to use the principles in a general way, when there are not contraindications from other sources, in an attempt to build protocols, look at protocols, and analyze them from the viewpoint of what they will do to the resistant subtypes.

Q: (Dr. Simon) If you have an agent A that works 70% of the time and an agent B that works 40% of the time, do you still obtain ABAB as an optimal schedule?

A: I should be quite exact and say that this is only for agents that are equally effective. When you have agents of differing effectiveness, it's very difficult to know what to do, because there are a large number of factors which can influence their differing effectiveness and thus would influence your decisions about which way to go. For example, the lower effectiveness may be due to smaller log kills or to differences in mutation rates; you would have two different decisions. Now, unless you know really what is the cause of difference in effectiveness, you don't know what to do. You just don't have good enough data from human neoplasms to use this model to actually produce protocols. However, I think what you can get from this kind of model is some idea about what protocols you should be trying to do—to control resistance. So, in cases such as the one I outlined, it seems to be a clear case that if you combine as many active agents as possible, you should get a better effect on resistance; you should get better control of it. It doesn't say that you're bound to get better control of it, because cross-resistance is quite a common phenomenon, but at least you shouldn't do any worse.

Q: (Dr. Don Young) I have a question in regard to sensitivity and selective response to treatment. Many of us have clonogenic stem cell assay laboratories set up, and I don't think very many people know what to do with the results. From a theoretical standpoint, and also from the standpoint of using that type of data in a predictive theoretical capacity, do you see any way in which these clonogenic stem cell assay results can be put to valuable clinical use?

A: What I would really like to know more about the stem cell assays is if in fact they always clone stem cells. I think there are viable tumor cells that one gets in the assay which might not necessarily be stem cells but which, nevertheless, are capable of some proliferation. If we can actually be sure that we're cloning stem cells then we should be able to use this in a predictive way by looking at the results when the cells are treated with a drug. We have to be very careful of the amount of drug that's given, careful that it corresponds with clinical levels of the drug that the tumor will be exposed to in vivo.

Q: (Dr. Harvey Schipper, Winnipeg) The model has been very stimulating for a long time, and not just to Canadians. There are, however, a fair amount of experimental data from Josh Fidler and others suggesting that one of the fundamental assumptions may not hold: that is, the constancy of the mutation rate. The suggestion has been made that as you reduce the size of the tumor, the mutation rate, or at least the rate of resistance induction, whatever that mechanism may be, goes up. Have you been able to look at that hypothesis? Because it seems to me it implies a different therapeutic strategy: rather than everything up front, this strategy suggests that the "late hit" may be a more appropriate approach. In the final analysis, I suppose, the question comes down to this: Is the strategy ever biologically testable in humans?

A: The first question is whether the mutation rate evolves. I think that in many systems it's well shown that it does evolve over time and tends to increase, so I would agree with you. What implications does this have for this model? Our original calculations would show that it just emphasizes our previous results. It seems to provide more emphasis for destroying cells as quickly as possible, because the longer they're around the more likely they are to develop resistance. Whether it's ever going to be clinically testable is a difficult question for me to answer. I think that, generally, hypotheses in any kind of human disease are difficult to test clinically.

Q: (Dr. Robert Wittes, NCI) One of the striking things that has been observed clinically over the last few years in many common cancers is the failure of alternating putatively non-cross-resistant combinations to increase response rates or complete response rates significantly over what you would get if you used one combination until failure and then the other. That, plus the fact that the responses in many of the diseases that we treat are partial rather than complete, suggests that the prevalence of resistant cells in a clinically evident tumor is vastly higher than it is in experimental tumors that you've used to test the mathematical model. Do you have any comment about what the mathematical models would yield if you assume very high values for the mutation rates? This really relates to the preceding question about the applicability of the model to the clinical situation.

A: Well, very high mutation rates give you exactly what you think they give you—very resistant tumors. A number of mechanisms have been proposed for large levels of cross-resistance in in vitro lines, which may apply to human tumors as well. I think that it does appear that cross-resistance, or if it's not resistance, some other phenomenon of human tumors, is prevalent in some and not in others. What it is, I don't really know, but if it is cross-resistance, it does suggest perhaps some hope for the future. One could hope that this cross-resistant mechanism is perhaps single, because it seems to apply to such a large variety of different agents. Therefore, specific steps could be taken to overcome this mechanism. It's possible that once you overcome it in one line you would have something that would be useful in a large number of different tumor systems.

(Dr. Wittes) I think that what we know about drug resistance in cell lines suggests that there are many mechanisms. It seems unlikely that there are going to be unitary ways to get rid of them. Even within drug classes that's true. Just looking at methotrexate resistance, there are any number of ways that cells can evolve methotrexate resistance.

A: I think that's true, but how prevalent are these different mechanisms in human tumors? I think it's likely there must be some very common phenomenon that causes resistance to such a wide variety of agents.

Q: (Dr. John Kovach, Mayo Clinic) I just want to make one comment. One of the assumptions that you make but don't list is that there's uniform exposure of tumor cells to drug. While in leukemias this may be true, in solid tumors it is not. In fact, the resistance one is seeing may not be at the cellular level, at least not at the genetic level; it may simply be a lack of access to the drug. Do you have any information on that, about penetration of drug into tumor? At least that's an avenue that I think should be explored.

A: I think it's a reasonable concern. I don't have any particular information on it that other people wouldn't have. I just stress at this point that this is meant to be a very minimal model that looks at one particular problem associated with the therapy of tumors. Of course, many other problems exist.

Response Surface Methodology and the Design of Clinical Trials for the Evaluation of Cancer Chemotherapy

Walter H. Carter, Jr[1,*]

I would like to address the problem associated with the development of optimal treatments in cancer therapy. This discussion will be restricted to the consideration of quantitative treatment levels, eg, dose of chemotherapy, dose of radiation therapy, time interval between doses of two treatment modalities, etc. Thus, the determination of levels of the treatment variables associated with minimum risk of failure for the treated individual is being considered. The current approach utilizing phase I (determination of toxicity), phase II (determination of activity), and phase III (comparative trial) studies does not generally result in the estimation of the optimal use of quantitative treatment modalities.

Imagine a phase III trial in which a new drug is being considered for use in combination with the standard therapy. Patients will be randomized to either the standard therapy or the new combination made up of the standard plus the new agent at a fixed dose and schedule. After an appropriate number of patients have been randomized to the study and an appropriate number of failures, (eg, deaths or tumor progression) have occurred, the data will be analyzed. The point of the analysis is to determine if there is a difference in the time to tumor progression or survival distributions associated with the two treatment arms. If there is, and the new treatment is superior, it can be concluded that, at the dosages used, the new treatment results in longer time to response than the standard therapy. However, it can not be concluded that the combination is being used optimally. As a result of the experimental design, there is no information from the study that can be used to determine the optimal levels of treatment. Suppose, as is often the case, that there is no statistically significant difference between the two treatment groups. The question immediately arises as to whether this is really the case or whether it is due to the new drug being used in a suboptimal way. Again, as a result of the experimental design, it is not possible to answer this question from the data collected as a result of the study.

I would like to discuss, via an example, a different, perhaps more informative approach to the problem of treatment evaluation.

The survival data given in table 1 result from an experiment in which 10^6 L1210 leukemia cells were injected sc in female $B6D2F_1$ mice on Day 0. Treatment began on Day 7. Dacarbazine (DTIC) was administered daily for 8 days and razoxane (ICRF-159) was given on Days 7, 11, and 15. It is important to note that the experiment involves a relatively large number of different treatment groups with not necessarily a large number of animals per treatment group. This experimental plan is one from the class of factorial experimental designs. As a result of this design, Cox's proportional hazards model can be used to approximate the relationship between the natural logarithm of the relative hazard and the levels of treatment as

$$\ell n\left(\frac{\lambda(t)}{\lambda_0(t)}\right) = \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2$$

where

X_1 = dose of DTIC (mg/kg)
X_2 = dose of ICRF-159 (mg/kg)
$\beta_1, \beta_2, \beta_{11}, \beta_{22}$, and β_{12} are unknown constants which parameterize the model.

The quadratic model is used to account for the toxicity associated with treatment. As a result of the properties of these drugs, the relative risk is likely to decrease with increasing treatment levels until a minimum is achieved and, due to their toxicity, increase as dosage levels are increased beyond this point.

The estimation of the parameters and verification of the model are key steps in the analysis, but that is not the point of this presentation. Suffice it to say that the unknown parameters can be estimated from the data, and it can be shown that there is close agreement between the observed and modeled data. Once the estimated dose-response relationship has been shown to be in close agreement with the observed data, the following information

[1]Departments of Biostatistics, Medicine, and Pharmacology and Toxicology, Medical College of Virginia, Richmond.

[*] *Reprint requests to:* Dr. Walter H. Carter, Jr, Departments of Biostatistics, Medicine, and Pharmacology and Toxicology, Medical College of Virginia, 1101 E Marshall St, Richmond, VA 23298.

TABLE 1.—DTIC + ICRF-159 combination experiment*

Treatment levels (mg/kg)		
DTIC	ICRF-159	Survival times in days†
0.0	0.0	10 (2), 11 (4), 12 (2)
60.0	0.0	11 (3), 12 (4), 14
90.0	0.0	11 (3), 12 (2), 13 (2), 14
135.0	0.0	10, 11 (5), 13 (2)
0.0	267.0	13 (3), 14, 18, 19, 20, 24
0.0	400.0	13, 16, 18, 24, 25 (3), 26
0.0	600.0	12, 13, 15 (3), 18 (2), 19
60.0	267.0	14, 22, 24, 25, 26 (2), 27
60.0	400.0	12, 13, 14, 19, 27 (2), 28, 29
60.0	600.0	11 (2), 12 (4), 17, 27
90.0	267.0	18, 19, 21, 22, 23 (2), 25, 28
90.0	400.0	15, 17, 26, 27 (2), 28, 29 (2)
90.0	600.0	13 (2), 14, 15, 16 (2), 17
135.0	267.0	10, 12, 15, 17, 18, 21, 22, 24
135.0	400.0	10, 15, 17 (2), 18, 19, 20 (2)
135.0	600.0	11 (2), 15, 16 (2), 17, 19, 22

*From Carter WH, Jr, Wampler GL, Stablein DM, et al. Drug activity and therapeutic synergism in cancer treatment. Cancer Res 42:2963-2971, 1982.

†Nos. in parentheses = No. of animals dying on the given day.

can be obtained from an exploration of the response surface:

1. The dose-response surface can be plotted. Figure 1 is a plot of the dose-response surface in which the darker shadings indicate regions of improved treatment, ie, lower risk or longer survival. If better treatments could be found outside the region of the experiment, darker shadings would occur at the periphery of the treatment space.
2. The estimated optimal treatment can be determined by finding the levels of X_1 and X_2 associated with minimum

$$\ell n\left(\frac{\lambda(t)}{\lambda_0(t)}\right),$$

or equivalently, maximal survival time.

For this experiment these levels are:

DTIC = 66.7 mg/kg

ICRF-159 = 370.1 mg/kg

for the given schedule of administration. In performing the optimization, the actual treatment space defined by the experiment was explored. Thus, if the optimal treatment levels were found to equal the largest dose used, this would indicate the possibility that the true optimum would be outside the region of the experiment. This would be indicated also by the plot of the response surface. In clinical applications, it is likely that the unconstrained optimal treatment would be outside the treatment space defined by a clinical trial and would be associated with unacceptable toxicity. Consequently, it is important to note that the methodology to accommodate optimization in the presence of constraints has been developed.

3. The confidence region about the location of the optimal treatment combination can be determined. Figure 2 is a plot of the 95% confidence region about the location of the optimal treatment. Note

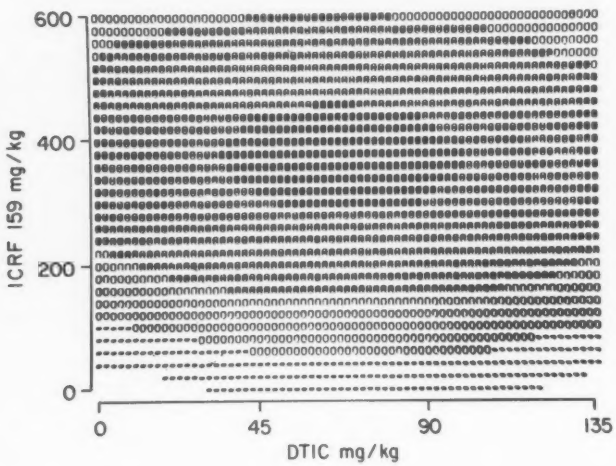


FIGURE 1.—Plot of dose-response surface in which the darker shadings indicate regions of improved treatment, ie, lower risk or longer survival. (*From* Carter WH, Jr, Wampler GL, Stablein DM, et al. Drug activity and therapeutic synergism in cancer treatment. Cancer Res 42:2963-2971, 1982.)

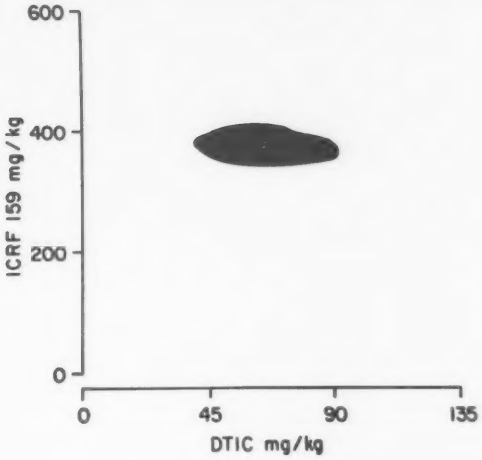


FIGURE 2.—Plot of the 95% confidence region about the location of the optimal treatment. (*From* Carter WH, Jr, Wampler GL, Stablein DM, et al. Drug activity and therapeutic synergism in cancer treatment. Cancer Res 42:2963-2971, 1982.)

that it is closed and located in the interior of the treatment space. Notice further that it does not make contact with either drug axis. Thus, it can be concluded with 95% confidence that the optimal treatment includes nonzero doses of either agent; ie, better treatment, as indicated by longer survival, can be obtained with the combination than with either agent at its individual optimal dose. When this is the case, therapeutic synergism can be claimed. Further, a statement has been made about the effect of the combination at its optimal use compared to the optimal use of each of the single agents. This is not possible with the current methodology.

The Mid-Atlantic Oncology Program (MAOP) is currently running a trial evaluating the interaction of dosage and scheduling of methotrexate (MTX) with 5-FU in combination with cyclophosphamide for the treatment of metastatic breast cancer. The design consists of 24 different treatment combinations which are obtained by crossing four levels of MTX (40, 80, 120, and 200 mg/m^2) in a factorial manner with six time intervals between MTX and 5-FU (0, 2, 4, 8, 16, and 24 hrs). The recommended sample size for this study is 312 patients, ie, 13 patients/treatment combination. It is important to realize that the purpose of this design is to gain information about the underlying dose-response relationship as opposed to making comparisons between individual treatment groups. Consequently, the number of treatment groups is large while the number of patients randomized to each treatment group is small. This sample size proposal is based on simulation studies of potential clinical situations. These simulations: (*a*) assumed a response to therapy; (*b*) generated exponential failure times from a 6 × 6 factorial with nine replications; (*c*) followed all subjects to failure; (*d*) estimated the response surface and its optimum using a proportional hazards model; (*e*) placed a confidence region about the optimal treatment levels; and (*f*) repeated the process 300 times and determined the number of times the control treatment (eg, standard dose of drug and simultaneous administration) was excluded from the confidence region.

In the simulation studies, situations were modeled which could occur in practice. For example, it was assumed that the optimum combination was in the center of the treatment space with a relative hazard of 0.55. Each of the single agents was less efficacious than the combination and was equally effective at its individual optimum, having a relative hazard of 0.72. With 312 patients, 90% of such studies resulted in the conclusion of improved combination modality therapy relative to the standard therapy. Another situation studied proposed that one dimension (MTX) had essentially no impact while the other (timing) was such that a relative hazard of 0.66 is achieved at its optimum located near the center of the range of possible values. In 74% of these studies, it was possible to exclude control treatment from the confidence region about the optimal levels of treatment.

Although there are no results yet available from this study, it does serve to illustrate the relationship between experimental design and the information available as a result of a response surface analysis of the study data. From the MAOP study, it will be possible to: (*a*) compare the effects of short versus long time intervals between MTX and 5-FU; (*b*) compare low versus high doses of MTX; (*c*) evaluate the interaction of the time interval between MTX and 5-FU and the level of MTX; and (*d*) estimate and explore the treatment-response relationship. The first and second possibility given above could be accomplished as a result of the usual phase III trial, whereas the third and fourth could not.

It should be noted that there are other designs, not factorial in nature, which permit this type of analysis and do not require as many different treatment combinations. Their use should also be explored. Current research in this area is involved with the development of methods which result in a reduction of the size of the confidence region about the optimal treatment levels. If successful, this would have the effect of reducing the number of patients required and thus make the response surface methodology even more applicable to clinical research.

DISCUSSION

Q: (Dr. Cyrus Mehta, Harvard) It's true that this type of a design would yield a lot more information than the conventional clinical trials design where you have only a single level of doses. We have, however, found one limitation in our studies with the Eastern Cooperative Oncology Group. It is difficult for an institution to handle more than four or five treatment regimens at any given time. So that to have the possibility that you could randomize to one of 24 different treatment regimens in a multi-institutional clinical trial might lead to practical problems. I wonder if you will face any of these problems with the study that you are talking about, or is that a single-institution study?

A: (Dr. Carter) This is a cooperative trial. So far we've not faced those kinds of problems.

Q: (Dr. James Anderson, Harvard) My concerns relate to the model specification. In the study that has been proposed, it's not clear to me whether you're going to analyze using 23 indicator variables to specify the treatment groups, or to specify some simple relationship among the treatment variables. The other question I have is: on top of this variability in treatment, you have variability in patients, and I wonder how you propose to take into account the wide deviation in performance status and other prognostic factors that the patients come to the trial with.

A: To answer the second question, we will use the Proportional Hazards Model to adjust for prognostic factors. In all clinical trials you have variability from patient to patient, so that's not something new that is introduced as a result of this type of design.

With regard to your first question, the model that we would fit for that study would probably include quadratic terms for the dose of methotrexate and either linear or quadratic terms for the time interval between the two. We would go through a model-building procedure there, and not account for each individual treatment group with an indicator variable.

Q: (Dr. Robert Wittes, NCI) I just wanted to follow up on the question about prognostic factors. You're certainly right that problems with imbalances of prognostic factors aren't unique to designs like this. But what is relatively unique is the small number of patients per cell that you're going to end up with, and it would seem that the probability for significant imbalance of important factors would be larger. Can you explain to a nonstatistician how the model deals with this? It would seem very difficult to get interpretable results out of it, unless you have enormously large numbers of patients.

A: One way of getting at the answer is that the inferences that are being drawn are not being drawn from the individual treatment combinations—the 24 treatments. We're not trying to compare cells within the design. We're trying to get information about the underlying dose-response relationship between the timing and the dose levels. Potential prognostic imbalances among cells are adjusted for by including in the regression model terms for the effects of these prognostic variables.

Factorial Designs for Randomized Clinical Trials

David P. Byar* and Steven Piantadosi[1]

Although hundreds of research papers and several books have been devoted to methodology for randomized clinical trials, we are unaware of any works dealing specifically with the subject of factorial designs for these trials in any detail. Considering the great usefulness attributed to these designs in agricultural and industrial experimentation, we find it unusual that the use of factorial designs has not been considered more widely in designing randomized clinical trials. Undoubtedly, one reason is that those who organize trials are unfamiliar with factorial designs. Another important reason might be that factorial designs require that the various treatments under study be capable of simultaneous administration to the patients without reduction in dose. This consideration would clearly have made factorial designs unsuitable for many studies of combination chemotherapy. We were drawn to the subject of factorial designs because of recent interest in designing studies to test whether or not vitamins or other nutritional supplements or pharmacologic agents such as aspirin might reduce the incidence of cancer or cardiovascular disease. Such studies are often called "prevention trials" to distinguish them from clinical trials designed to compare two or more treatments in patients with known disease. Such trials may be particularly well-suited to factorial designs, because the agents studied should be almost completely nontoxic so that they may be given to supposedly healthy subjects of whom only a relatively small fraction will be expected to develop the diseases under study. This means that it is actually feasible to give two, three, or even more agents simultaneously to some subjects, in sharp contrast to the situation that applies with drugs used to treat diseases already in existence, such as giving multiple chemotherapeutic agents to patients with cancer. In the latter case, overlapping toxic effects commonly require that the dosages of the individual drugs be reduced when they are used in combination, making factorial designs inappropriate.

We make no claim to originality of the concepts discussed in this article; our purpose is simply to introduce some ideas about factorial designs and consider what advantages they may have in medical applications. We will touch briefly on such important concerns as multiple comparisons, interactions, and statistical power.

Basic Idea of Factorial Designs

The simplest possible factorial design is shown at the top of table 1. This is called a 2^2 or 2×2 complete balanced factorial design. There are two treatments, A and B, and patients are assigned with equal probability to the four treatment groups: neither treatment, A or B alone, or both together. To illustrate how the idea of factorial designs generalizes, a design for studying three treat-

Table 1. Two complete balanced factorial designs

2^2 or 2x2:

A \ B	−	+	
−	n	n	2n
+	n	n	2n
	2n	2n	4n

2^3 or 2x2x2:

A	B	C	
−	−	−	n
−	−	+	n
−	+	−	n
−	+	+	n
+	−	−	n
+	−	+	n
+	+	−	n
+	+	+	n

[1]Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, Bethesda, MD.

* *Reprint requests to:* Dr. David P. Byar, Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, Blair Bldg, Rm 5A05, National Institutes of Health, Bethesda, MD 20205.

ments, called a 2^3 or $2 \times 2 \times 2$ factorial design, is shown at the bottom of table 1. There are eight treatment groups. One group receives none of the three treatments, three groups receive one treatment each, three groups receive two treatments each, and one group receives all three treatments. In both the 2^2 and 2^3 designs we note that exactly one-half of the patients receive each of the treatments under study and it is this fact which confers the great efficiency on factorial designs. We will now turn to the analysis of such designs, but we will note in advance that whether we are estimating main effects or various orders of interactions, we always use all of the data from the experiment, simply combining them in different ways.

In table 2 we show estimation of the main effects and interaction for a simple 2×2 factorial design. We imagine that we estimate $\bar{X}$, the mean value of some normally distributed response in each of the four treatment groups. Our notational convention will be that the subscript on $\bar{X}$ indicates the treatment used in that group; the absence of a subscript indicates the absence of both treatments. For simplicity we will begin by considering estimation of the main effect of Treatment A. The design offers us two estimates of the main effect of A, $\bar{X}_a - \bar{X}$ and $\bar{X}_{ab} - \bar{X}_b$, the first estimate being the effect of A in the absence of B and the second being the effect of A in the presence of B. We therefore average the two estimates as indicated in table 2. Estimation of the main effect of B is completely analogous.

We next turn to the estimation of the interaction between A and B. A nonzero interaction between A and B means that the effect of A is different depending on the presence or absence of B. We are therefore interested to see if our two estimates of the main effect of A, namely, $\bar{X}_a - \bar{X}$ and $\bar{X}_{ab} - \bar{X}_b$, are different, so we subtract them, using their difference as an estimate of the interaction. Since we only have one estimate of the interaction, we do not divide by two as we did when estimating the main effects of A and B. Thus, even though we use all of the data when estimating the interaction, the variance of this estimate is four times as great as that for the main effects, implying that our power to detect significant interactions is less than that for detecting significant main effects. We also note that the interaction of A with B is logically equivalent to the interaction of B with A, and in fact the equations given at the bottom of table 2 are algebraically identical.

Table 2. Estimation of main effects and interation for a 2x2 factorial trial with n patients in each cell

A \ B	no	yes
no	$\bar{X}$	$\bar{X}_b$
yes	$\bar{X}_a$	$\bar{X}_{ab}$

$$\text{Main Effect of A} = \frac{(\bar{X}_a - \bar{X}) + (\bar{X}_{ab} - \bar{X}_b)}{2}$$

$$\text{Main Effect of B} = \frac{(\bar{X}_b - \bar{X}) + (\bar{X}_{ab} - \bar{X}_a)}{2}$$

$$\text{Interaction}^* \text{ of A with B} = (\bar{X}_a - \bar{X}) - (\bar{X}_{ab} - \bar{X}_b)$$

$$\text{Interaction}^* \text{ of B with A} = (\bar{X}_b - \bar{X}) - (\bar{X}_{ab} - \bar{X}_a)$$

*In Additive Model

Although we have explained the estimation of main effects before explaining the estimation of interaction, in practice one should test for significant two-way interactions before computing the main effects, because if significant interactions are present, then the usual estimates of the main effects are not of primary interest, unless we choose to regard the main effect of Treatment A, let us say, as an average over a population of subjects one-half of whom were treated by B and one-half of whom were not.

The analysis of a 2^3 design is shown in table 3. At the bottom of the table we have shown how to estimate the main effect of A. We note that the design allows four estimates of the main effect of A. The first is in the absence of Treatments B and C, the second is in the presence of B but not C, the third is in the presence of C but not B, and the fourth is in the presence of both B and C. We therefore average these four estimates. Estimation of the main effects of Treatments B and C proceeds analogously.

We have chosen only one of the three two-way interactions for the sake of illustration. We note that unlike the 2×2 design, the 2^3 design gives us two estimates of the two-way AB interaction: one in the absence of Treatment C and one in its presence. Therefore, we average these two separate estimates. The procedure is completely analogous for the AC and BC interactions. We finally come to the three-way interaction between Treatments A, B, and C. What do we mean by a three-way interaction? There are three entirely equivalent interpretations, just as there were two equivalent interpretations for the two-way interactions. These are that the AB interaction depends on the presence or absence of Treatment C, that the AC interaction depends on the presence or absence of Treatment B, or that the BC interaction depends on the presence or absence of Treatment A. The 2^3 design allows

Table 3. Estimation of main effects and interactions for a 2x2x2 trial with n patients in each cell

A	B	C	Results
-	-	-	$\bar{X}$
-	-	+	$\bar{X}_c$
-	+	-	$\bar{X}_b$
-	+	+	$\bar{X}_{bc}$
+	-	-	$\bar{X}_a$
+	-	+	$\bar{X}_{ac}$
+	+	-	$\bar{X}_{ab}$
+	+	+	$\bar{X}_{abc}$

Main Effect of A:

$$\frac{1}{4}\left[(\bar{X}_a-\bar{X}) + (\bar{X}_{ab}-\bar{X}_b) + (\bar{X}_{ac}-\bar{X}_c) + (\bar{X}_{abc}-\bar{X}_{bc})\right]$$

AB Interaction:

$$\frac{1}{2}\left\{\left[(\bar{X}_a-\bar{X}) - (\bar{X}_{ab}-\bar{X}_b)\right] + \left[(\bar{X}_{ac}-\bar{X}_c) - (\bar{X}_{abc}-\bar{X}_{bc})\right]\right\}$$

ABC Interaction:

$$\left[(\bar{X}_a-\bar{X}) - (\bar{X}_{ab}-\bar{X}_b)\right] - \left[(\bar{X}_{ac}-\bar{X}_c) - (\bar{X}_{abc}-\bar{X}_{bc})\right]$$

us a single estimate of the three-way interaction, and the algebraic form given at the bottom of table 3 is most easily understood in relation to the first interpretation just given. Note that the expression inside the square brackets to the left represents the AB interaction in the absence of C, whereas that within the square brackets to the right represents the AB interaction in the presence of C. Thus, their difference is an estimate of the ABC interaction. As before, we note that the precision of these estimates varies markedly. Estimates of main effects have variances four times smaller than two-way interactions and 16 times smaller than the three-way interaction.

The practical question of how many treatments or factors may be reasonably studied at once may arise. At least for the case where the response variable is normally distributed, theory is of no help in answering this question provided that 2^p (where p is the number of factors under consideration) is less than or equal to the total patients available for study. However, common sense suggests that overly complex designs should probably be avoided and more attention devoted to identifying the really important questions and assessing the feasibility of executing the design in practice. However, if four or more factors are identified as deserving simultaneous study, then fractional factorial designs should be considered. The idea of a fractional factorial design is that we may study only some fraction of the full 2^p cells if we are willing to sacrifice estimation of certain higher order interactions. These designs are described in standard textbooks on experimental design (1–4).

Interpreting Interactions

Whenever the subject of factorial designs comes under discussion, you can usually count on someone to say something like the following: "Yes, those designs are very nice, but what if you have interactions?" It appears that the principal concern being expressed by this question is that the overall estimate of the treatment effect may somehow be diluted if we use a factorial design, whereas we would not have to worry about this problem if we studied each treatment separately. Consider for example a 2×2 design in which $\bar{X} = 100$, $\bar{X}_a = 50$, $\bar{X}_b = 50$, and $\bar{X}_{ab} = 50$. We note that in the absence of B, the effect of A is to decrease the response by 50 units, whereas it appears to have no effect in the presence of B. The overall estimate of the effect of A, if we ignore the interaction, would be 25, only one-half of the 50-unit change we expect from comparing $\bar{X}$ and $\bar{X}_a$. If the numbers used in this example had been obtained from a study with a large sample size, we would have detected a very significant AB interaction, and then rather than use the usual estimates of the main effects of A and B, we would report our finding concerning the interaction along with the estimated effects of A and B when used alone. In fact, the interpretation of this experiment would be that Treatments A and B were equally effective and that the combination of the two offered nothing over using either one alone. We have therefore learned something that we could not have learned had we not used a factorial design or some variant of it. Whether or not outcomes such as this are likely to occur in practice depends on the nature of the treatments studied and whether or not it is likely that either one will show any effect. If we think the probability is low that either A or B has an effect, then it is even less likely that they will show a significant interaction.

Another way to view the problem of interaction among the treatments is to imagine that one of the treatments is simply another stratification factor, such as age or severity of disease. It is commonplace in the analysis of clinical trials to find that a treatment does not appear to be equally effective among all subsets of patients, although only when the variation across subsets is very marked or the trial is very large do we have enough data to conclude that the variation exceeds what we might expect due to chance alone. Thus, if we were designing a trial where we feel that certain interactions are likely, based on previous experience or biological argument, then we might consider designing the trial considerably larger than would be needed to detect main effects to have power for studying the interactions. If we are not particularly interested in interactions, then we should not let concern about their possible existence dissuade us from using factorial designs any more than we would restrict trial entry criteria to a very narrow range of age or severity of disease for fear that the treatment effect would not be constant across a broader range. However, after a trial is completed and a physician must choose treatment for future patients, the analogy is no longer helpful because, although he cannot control the stratification factors, he can control the treatments he uses, and he will choose according to his understanding of the trial results. For example, if results with A and B together are significantly worse than for either alone, he will not choose to give both, but interactions of a magnitude plausible in medical settings are only likely to be detected if they are significant in very large trials.

When considering interactions from the point of view of biological interpretation, we should keep in mind that the concept of interaction is model-dependent. This means that the presence or absence of interactions may simply represent the scale of measurement. The estimation of second- and third-order interactions discussed in the previous section was based on the assumption of an additive model. For some problems, however, a multiplicative model is more appropriate. It should be remembered that complete absence of an interaction on an additive scale implies interaction on a multiplicative scale and vice versa. This point is illustrated with some hypothetical data in table 4. In the left part of the table we see a sample outcome representing no interaction on the addi-

Table 4. Hypothetical data illustrating the model dependence of interactions

		B	
		no	yes
A	no	10	20
	yes	30	40

No Interaction For Additive Model

		B	
		no	yes
A	no	10	20
	yes	30	60

No Interaction For Multiplicative Model

tive scale, since the effect of A is to increase the response by 20 units whether or not B is present. In the example to the right, however, the effect of Treatment A is to multiply the response by a factor of three whether Treatment B is present or not. Thus, while there is no interaction for an additive model in the example to the left, there is interaction for a multiplicative model, since Treatment A multiplies the response by a factor of three when B is absent but only by a factor of two when B is present. The reverse is true for the example to the right; there is no interaction on the multiplicative scale, but there is on the additive scale. In both of these examples, the effect of Treatment A is in the same direction whether B is present or absent. However, if there is a reversal of the effect of A when B is present compared to when it is absent, then we would say that there was a qualitative or crossover interaction. Interactions of this kind cannot be made to disappear by a simple change of scale.

Problem of Multiple Comparisons

It is sometimes suggested that one cannot really answer two or three questions for the price of one by means of factorial designs because we must take into account the fact that we are asking two or three questions rather than one, and therefore we should adjust the significance level at which we choose to work, say 0.05, in such a way that the overall probability of rejecting the null hypothesis when it is true is set at that nominal level. Although this approach can be justified from a frequentist point of view, it does not appear to have attracted much attention in standard text books on experimental design. Suppose that instead of comparing Treatments A and B in 2×2 factorial design, we had decided instead to perform two separate trials. Would we then have insisted that each of these trials should be analyzed using an α-level of 0.025 on the grounds that we wanted the probability of a type I error to be guaranteed at 0.05 over the two trials together? In fact, suppose that over the rest of our life we were going to be involved in 100 trials. Would we then feel compelled to adjust the α-levels for each trial in such a way that over our entire life we would reject the null when it was true only 5% of the time? Or suppose we were reading a scientific journal containing 20 articles. Do we mentally adjust the P values reported in those articles in such a way as to set the probability of a type I error for our entire reading session at a prespecified value? It is unlikely that most of us behave in this way, and this argument can be used to justify working at the usual α-levels rather than at adjusted ones. Perhaps the most important point to remember is that we should not really believe P values unless they are small, say < 0.01. Even more important than this "common-sensical" piece of advice is the observation that repetition of the results of the experiment by other investigators in different settings is likely to be a much more reliable guide to scientific truth than any P value, however small, adjusted or not.

P values are of course very useful to statisticians in distinguishing possibly real effects from random variation, but they may not be regarded as appropriate measures for assessing the strength of evidence in all settings. Perhaps the most appropriate use of P values according to the frequentist point of view is when we are interested in the properties of some testing procedure rather than in the outcome of any particular test. For example, we might want to set up a procedure for inspecting manufactured items in such a way that we would not too often reject items which were in fact not defective, but we might have little interest in the outcome of any particular test. This is a rather different situation from that of trying to assess the outcome of a single randomized trial which is unlikely ever to be repeated in exactly the same way. For those interested in thinking further about this subject, Cornfield (5) has provided some very interesting reflec-

tions on the limitations of *P* values.

Comparing Various Designs

Let us assume that we are genuinely interested in studying each of two treatments, A and B, and consider the advantages and disadvantages of various design options. Let us suppose (*a*) that we will perform significance tests for the main effects at $\alpha = 0.05$ (one-tailed and unadjusted for multiple comparisons) and (*b*) that the control group, designated by ϕ and representing the absence of either Treatment A or B or both, has a failure rate of .5 and that we wish to design a trial in such a way that we have $\geq 90\%$ power for failure rates of .3 for either Treatment A or B. When considering factorial trials we will deal both with the situation where there is no interaction between Treatments A and B on the additive scale, which means that the combination of Treatments A and B will have a failure rate of .1, and also with the situation where there is some interaction between Treatments A and B, so that their joint effect yields a failure rate of .2. The various designs under consideration are shown in table 5. First, we consider two two-arm trials, each with 120 patients on each of its two arms, for a total of 480 patients. This design will give us approximately 93% power under the alternatives specified above. The assessments of Treatments A and B are independent for this first option, but no information on possible interactions between the two is provided. The second option is a single three-arm trial with 120 patients on each arm. We still have a power of 93%, even though we have only used 75% as many patients. However, the assessments of the two treatments are not independent because they share the same control group, and there is no information on interactions. For option 3 we consider a 2 × 2 factorial design. In 3a we place 60 patients in each of the four treatment groups and obtain slightly greater power for testing each of the main effects (for reasons that need not concern us here), even though we are only using one-half as many patients as in option 1, that of two separate two-arm trials. It does appear that we have gotten two answers for the price of one. However, had we been concerned in advance that the interaction assumed above might exist in nature and had desired to maintain high power for testing main effects, we might have increased the sample size in each of the four arms by 50%, resulting in 90 patients/arm, for a total of 360 patients. In fact, if the interaction were absent the power would have risen to 99%, but in its presence we would still have had 92% power for testing the main effects.

Although our reason for increasing the number of patients in each cell from 60 to 90 was to maintain high power for detecting the main effects in the presence of dilution by the assumed interaction, it is interesting to note that the power for testing the assumed interaction would only be about 18%, emphasizing the point noted above that modest interactions will achieve significance only in very large studies. In fact, to achieve a power of 80% (or 90%) for detecting the interaction assumed in table 5, we would need 651 (or 871) patients in each of the four cells. Of course, if such a study were done and the interaction were significant, then with those much larger numbers the power would approach 100% for estimating the main effect of A in the absence of B. However, with only 90 patients/cell, a sample outcome corresponding to the assumed interaction in table 5 would not even approach significance, and therefore testing of the main effects would be appropriate.

These factorial designs in options 3a, 3b, and 3c provide information on the presence or absence of interactions and allow independent estimation of the treatment effects. The final option we consider is a single two-arm trial in which we compare the combination of Treatments A and B to a control group, placing 120 patients in each of the two groups. This trial requires the same total number of patients as the 2 × 2 factorial design in option 3a and has approximately the same power if we assume that only one of the two treatments works and therefore that the failure rate for the AB treatment is 0.3. However, after the trial we would not know which of the two treatments worked, or if both worked, or if both worked but only if given together. This is another way of saying that we would have no information on any possible interaction between the two treatments. If both treatments were active and additive, so that the AB group had a failure rate of .1, then the power would be greater.

When To Use Factorial Designs

Factorial designs should be considered whenever one is genuinely interested in studying more than one treatment. This is, however, almost always the case since there are many more unanswered questions about medical treatment than there are resources to answer them. To use a factorial design it is necessary that the two or more treatments can actually be given together without modification, that is to say that they are not known to interfere with each other and their toxic effects, if any, are not likely to attain unmanageable levels. As mentioned briefly in the introduction, substances such as vitamins and trace metals now being studied in prevention trials are likely to be good candidates for factorial designs. However, even in cancer treatment studies, factorial designs are sometimes appropriate. Staquet and Dalesio (6) cite a European Organization for Research on Treatment of Cancer adjuvant study in epidermoid carcinoma of the lung where 600 patients were studied in a 2 × 2 factorial trial evaluating both chemotherapy and immunotherapy. Another example of a therapeutic trial is one currently being conducted by the Brain Tumor Study Group in the United States. Here, two questions concerning the chemotherapeutic drugs BCNU and 5-FU are being an-

Table 5. Comparing power for testing main effects for various designs

Assumptions: (1) one-tailed tests, alpha = 0.05
(2) Control (ϕ) = .5, A or B = .3
(3) AB = .1 (no interaction[1]) or
AB = .2 (interaction[2])

Designs	Schema	Total Patients	Power	Comments
1. Two 2-arm trials	ϕ - 120 A - 120 ϕ - 120 B - 120	480	93%	-Independent -Largest sample size -No information on interactions
2. One 3-arm	ϕ - 120 A - 120 B - 120	360	93%	-Not independent -No information on interactions
3a. 2x2 factorial no interaction	ϕ - 60 A - 60 B - 60 AB - 60	240	96%	
3b. 2x2 factorial no interaction	ϕ - 90 A - 90 B - 90	360	99%	-Independent -Information on interactions
3c. 2x2 factorial interaction	AB - 90	360	92%	
4. Single 2-arm	ϕ - 120 AB - 120	240	93%	-Which works? -No information on interactions

[1]No interaction:

	no B	B
no A	.5	.3
A	.3	.1

[2] Interaction:

	no B	B
no A	.5	.3
A	.3	.2

swered simultaneously in the 2 × 2 factorial trial. (*a*) Is intracarotid carmustine (BCNU) better than iv BCNU? (*b*) Is 5-FU plus BCNU better than BCNU alone? Accordingly, investigators plan to accrue 600 patients in each of four arms: iv BCNU without 5-FU, intracarotid BCNU without 5-FU, 5-FU plus iv BCNU, and 5-FU plus intracarotid BCNU. In this trial there is no untreated control group; yet, the principles are exactly the same, since we may regard the route of administration of BCNU as factor A, where we may take the absence of A as iv BCNU, for example, and its presence as intracarotid BCNU. Factor B may be regarded as the presence or absence of 5-FU. Both of these cancer trials illustrate a principle that some researchers have advised for the medical application of factorial designs, namely, that the mechanism of action of the factors or treatment under study be different.

It has been suggested that factorial designs would be most appropriate when serious interactions are not expected, and this is one reason for suggesting that it is desirable that the mechanisms of actions be different. However, if one is particularly interested in studying interactions, or one expects that they are likely to exist, then factorial designs provide an efficient, logical, and systematic approach for studying them. But because of lower power for detecting significant interactions relative to that for detecting main effects, one should increase the sample size accordingly in these situations.

REFERENCES

1. COCHRAN WG, and COX GM. Experimental Designs, 2nd ed. New York, John Wiley & Sons, Inc, 1957.
2. COX DR. Planning of Experiments. New York, John Wiley & Sons, Inc, 1958.
3. FISHER RA. The Design of Experiments, 8th ed. New York, Hafner, 1960.
4. KEMPTHORNE O. The Design and Analysis of Experiments. New York,

John Wiley & Sons, Inc, 1952.
5. CORNFIELD J. Sequential trials, sequential analysis and the likelihood principle. *In* The American Statistician, 1966, pp 18–23.
6. STAQUET M, and DALESIO O. Design of phase III trials. *In* Cancer Clinical Trials: Methods and Practice (Buyse ME, Staquet M, and Sylvester RJ, eds). New York, Oxford University Press, 1984.

DISCUSSION

Q: (Dr. Richard Simon, NCI) To add some concreteness to the discussion of the issue of factorial designs, consider the following example. Suppose you're planning a trial in stage 2 breast cancer to compare an adjuvant chemotherapy to no chemotherapy. Someone says, you could subrandomize patients for whether or not to receive postoperative radiotherapy and you could do that four-arm trial with the same number of patients as the two-arm trial. You could collapse the arms, which is really what the factorial design means. Do you agree with that?

A: (Dr. Byar) If you have no reason to believe that the two are likely to interact, and if you are not one of those people who likes to correct the *P* values, the answer would be "yes." If you do want to correct the *P* values, you will not have to do much; you just change the alpha level from .05 to .025, which will cause you to have a somewhat larger sample size, but nothing like that required for conducting two separate trials with power equal to that for the factorial design. Separate trials would provide no information about using adjuvant therapy and x-ray together and would require about twice as many patients. So I would find the factorial design attractive. This is a good example of two treatments that presumably work by different mechanisms.

Q: (Dr. Susan Ellenberg, NCI) One wrinkle on the idea of factorial designs that I've heard about recently is the idea of having one main question and several other ancillary questions that may require fewer patients. All patients would be randomized between Treatment "a" and Treatment "b," and then one subset of institutions would have a second randomization of some ancillary issue, and other institutions might subrandomize to study another ancillary issue. So you would have a couple of factorial things going on at the same time. What would you think about that kind of design?

A: I would probably try to persuade all institutions to collaborate in a $2 \times 2 \times 2$ design. The design you mention would be acceptable only if each institution has sufficient patients for adequate power, but if they do not, then it is an illusion to imagine that this design properly answers several questions.

Tree-Structured Survival Analysis[1]

Louis Gordon[2] and Richard A. Olshen[3,*]

In this note, tree-structured recursive partitioning schemes for classification, probability class estimation, and regression are adapted to cover censored survival analysis. The only assumptions required are those which guarantee identifiability of conditional distributions of lifetime given covariates. Thus, the techniques are applicable to more general situations than are those of the famous semi-parametric model of Cox. [Cancer Treat Rep 69:1065–1068, 1985]

Background

Recursive partitioning schemes for classification have proved useful in a variety of medical (and other) settings. Goldman et al (1) describe an application to the diagnosis of myocardial infarction. Olshen et al (2) study prognosis following myocardial infarction. Dillman and Koziol (3) apply partitioning methods to the diagnosis of cancer. See also Chapter 6 of *Classification and Regression Trees* by Breiman et al (4). We use the acronym Cart both for the cited book and for a generic binary decision tree used as an estimator in a classification, class probability estimation, or regression problem.

Notation and Motivation

The following discussion serves to introduce some notation, and to formalize the problem under consideration. Consider a $(q + 1)$ dimensional random vector $(\underset{\sim}{X},Y)$ having joint distribution $F(x,y)$, in which $\underset{\sim}{X}$ consists of q explanatory variables, and Y is a univariate response. We are interested in

$$E\left\{g(Y)|\ \underset{\sim}{X}\right\} = h(\underset{\sim}{X}) = \int g(y)F(dy|\ \underset{\sim}{X}),$$

the conditional expectation of g(Y) given the explanatory information in $\underset{\sim}{X}$. As usual, we refer to $h(\cdot)$ as the regression function of g(Y) on $\underset{\sim}{X}$.

Unfortunately, we do not know $h(\cdot)$, and wish to estimate it using a training sample consisting of n independent observations $(\underset{\sim}{X}_1,Y_1), \ldots, (\underset{\sim}{X}_n,Y_n)$. These observations are summarized by the joint empirical distribution function $\hat{F}_n(x,y)$. Were the random vector $\underset{\sim}{X}$ discrete, the estimate $\hat{h}_n(\underset{\sim}{X})$ would eventually consist of averages of response variables corresponding to distinct values of the explanatory variables $\underset{\sim}{X}$. Because the explanatory variables might not be discrete, we may imagine a Cart estimate of h to be derived by choosing a simple function s_n to discretize $\underset{\sim}{X}$, and then using

$$\int g(y)\hat{F}(dy|\ s_n(\underset{\sim}{X}))$$

to approximate

$$E\left\{g(Y)|\ \underset{\sim}{X}\right\}.$$

In reality, s_n is chosen from the data, and so there is work to be done in moving from the formalism of conditional expectation given discretized explanatory information to the certainty of mathematical proof.

The preceding heuristic underlies very general results regarding consistency of tree-structured methods presented in Gordon and Olshen (5–7) and in Cart, Chapter 12. In addition, the heuristic serves to suggest close connections with the mathematical theory of differentiation of integrals. See, for example, Shilov and Gurevich (8, p 213).

The objective of Cart is to use a sample to construct a binary decision tree. The preceding discussion indicates that the tree determines a partition which is used to

[1]Supported by Public Health Service grants HL-17682 (National Heart, Lung, and Blood Institute) and CA-26666 (National Cancer Institute) from the National Institutes of Health, Department of Health and Human Services, and by National Science Foundation grant DMS83-01338.

[2]Department of Mathematics, University of Southern California, Los Angeles.

[3]Department of Mathematics, Laboratory for Mathematics and Statistics, and Cancer Center, University of California-San Diego, La Jolla.

Reprint requests to: Richard A. Olshen, PhD, Department of Mathematics C-012, Laboratory for Mathematics and Statistics, and Cancer Center, University of California-San Diego, La Jolla, CA 92093.

approximate a (Radon-Nikodym) derivative. It is for this reason that Cart methods are applicable to problems of density estimation, discrimination, and regression.

The same idea of expressing an estimator in terms of the empirical cumulative distribution appears in Breslow and Crowley's (9) analysis of the large sample properties of the Kaplan-Meier product limit estimator of survivorship. For simplicity of exposition, we consider the situation in which death and censoring have absolutely continuous distributions with respect to Lebesgue measure. In order to avoid the problems of identifiability pointed out in Tsiatis (10), we make the fundamental assumption that death time and censoring time are conditionally independent, given the q prognostic explanatory variables $\underset{\sim}{X}$.

Formally, we consider the data at hand to be n independent observations of triples $(T \wedge C, I(T<C), \underset{\sim}{X})$, in which T is interpreted as time of death, C as time of censoring, and $T \wedge C$ denotes the minimum of T and C; $I(T<C)$ is the indicator that death occurs before censoring. As was stated, assume that T and C are conditionally independent given $\underset{\sim}{X}$, and that they have absolutely continuous conditional distributions $F_T(t|\underset{\sim}{X})$ and $F_C(s|\underset{\sim}{X})$. We require also that $P(C>t|T = t, \underset{\sim}{X} = \underset{\sim}{x}) > 0$ for all positive t and all $\underset{\sim}{x}$. A conditional form of Breslow and Crowley's discussion [(9), pp 445–446] and the methods of Gordon and Olshen (6,7) show that if our assumptions are valid, then with appropriate constraints on the adaptiveness of the algorithm Cart is applicable to the censored data problem with covariates. Note that the ultimate estimator consists of a partition of the explanatory variable feature space and a set of Kaplan-Meier estimates, one for each subset in the partition.

Implementation

The practical implementation of algorithms in our context depends on choices of splitting criteria, pruning criteria, and terminal decision rule. In addition, some form of cross-validation is necessary to estimate accuracy. In the next paragraphs, we sketch possible implementation of Cart-like procedures for estimating conditional survivorship. In that context, we illustrate a number of considerations which should govern any implementation.

Given any node in a tree $\mathcal{T}$, we compute the Kaplan-Meier estimate of the survival function within that node. Because we use the Kaplan-Meier estimator within nodes as we grow the tree, the natural splitting criteria involve measures of distance between estimated survival functions in candidate left and right daughter nodes. A convenient class of distance functions are the so-called L^p Wasserstein metrics, $p \geq 1$, which can be motivated as follows. First, suppose that the distributions in question are proper. If F_Z and F_W are two such distributions functions, then their L^p Wasserstein distance is the minimum of

$$\left[E\left\{|\, Z - W \,|^p\right\}\right]^{1/p},$$

computed for all Z, W defined on the same probability space, with Z distributed as F_Z and W as F_W. The L^p Wasserstein distance between empirical distributions figures in a recent multivariate generalization of the one-dimensional one- and two-sample Cramér-von Mises tests due to Ajtai et al (11). In fact, in one dimension, as is the case here, the L^p Wasserstein distance $d_p(F_Z,F_W)$ can be written explicitly as

$$\left[\int_0^1 |\, F_Z^{-1}(u) - F_W^{-1}(u) \,|^p du\right]^{1/p}, \qquad [1]$$

where $F_z^{-1}(u)$ is a consistently chosen $100u^{\text{th}}$ percentile of F_Z. If $p = 1$, then

$$d_p(F_Z, F_W) = \int_0^\infty |\, F_Z(v) - F_W(v) \,|\, dv.$$

Suppose now that F_Z or F_W is improper, so

$$\lim_{v \to \infty} F_Z(v) = m_Z \leq 1 \,;\; \lim_{v \to \infty} F_W(v) = m_W \leq 1.$$

If $m_Z = m_W$, then [1] applies with the upper limit of integration replaced by the common value of m. If, say, $m_Z < m_W$, then the Wasserstein distance between F_Z and F_W can be described informally thus. Let $L = F_W^{-1}(m_Z)$. If F_W has no mass at L, then let

$$\tilde{F}_W(v) = F_W(v) \text{ for } v \leq L, \text{ and } \tilde{F}_W(v) = m_Z \text{ if } v \geq L. \qquad [2]$$

Define $d_p(F_Z,F_W)$ to be $d_p(F_Z,\tilde{F}_W)$. If $m_Z < m_W$ and $\tilde{F}_W$ has an atom at L, then split the mass, if necessary, so that $\tilde{F}_W$ satisfies [2], and define $d_p(F_Z,F_W)$ as before. Finally, define the Wasserstein distance between two survival functions to be the distance between their corresponding distribution functions. The possible advantages of the cited Wasserstein metrics are: (*a*) they handle censoring simply and in an intuitively appealing fashion; and (*b*) by concentrating on "horizontal" distances between distributions, they focus naturally on tails of distributions and thus on "cures."

In the remainder we focus attention on the case $p = 2$, though we expect to consider other possibilities in the future. For any (possibly improper) survival function S, let δ_S be the survival function (possibly improper) that has mass at at most one finite point and that minimizes the Wasserstein distance to S. Finding δ_S is easy for any p since the function which needs to be minimized is convex. We take $d_2(S,\delta_S)$ as a measure of the variability of S. In

applications $d_2(S,\delta_S)$ will always be finite, since S's produced by the Kaplan-Meier procedure are step functions. (If S is proper, δ_S concentrates its mass at the expectation of S.) Because our splitting criteria are designed to force each node to more homogeneous as to survival function than its predecessors, measures of variability figure in the definitions of the criteria. Our first splitting criterion, then, is this. Given a node B of a tree, and a Kaplan-Meier estimate $\hat{S}^B$ computed from data at that node, cycle through the q axes of $\underset{\sim}{X}$ and split B as in Cart so as to maximize

$$\hat{P}(B)d_2(\hat{S}^B,\delta_{\hat{S}^B}) - \left[\hat{P}(L(B))d_2(\hat{S}^{L(B)},\delta_{\hat{S}^{L(B)}}) + \hat{P}(R(B))d_2(\hat{S}^{R(B)},\delta_{\hat{S}^{R(B)}})\right],$$

where $L(B)$ and $R(B)$ are candidate left and right daughter nodes of B, and

$$\hat{P}(B) = \#\left\{i \leqslant n : \underset{\sim}{X}_i \in B\right\}/n;$$

that is, $\hat{P}(B)$ is the empirical probability of the region of the feature space which corresponds to the node B. The splitting criterion maximizes the difference between weighted average variability in daughter nodes and variability in the parent node. If, for a given tree $\mathcal{T}$, $\tilde{\mathcal{T}}$ is its collection of terminal nodes, then a criterion for assessing how well $\mathcal{T}$ does in estimating the conditional distributions of T given $\underset{\sim}{X}$ is

$$\left[\sum_{B\in\tilde{\mathcal{T}}} \hat{P}(B)d_2(\hat{S}^B, \delta_{\hat{S}^B})\right] + \alpha\,|\,\tilde{\mathcal{T}}\,|\,, \qquad [3]$$

where $\alpha > 0$ is a constant and $|\tilde{\mathcal{T}}|$ is the number of terminal nodes of $\mathcal{T}$. The first term of [3] is a measure of how variable the terminal nodes are, while the second term is a penalty for lack of parsimony of the tree $\mathcal{T}$. The probability weights in the first term guarantee that more populous nodes figure more heavily in the overall assessment of the tree. A large tree is grown initially [Cart, p 63 and (10.29) on p 291]. In the construction there are built-in constraints to the cited splitting criterion. In fact, the tendency of tree-structured methods to be too adaptive (Cart, Section 11.8) is limited by the quantile splitting requirements of Gordon and Olshen (5–7), which are implemented by minimum node requirements (Cart, pp 63, 233). For each $\alpha > 0$ the large tree has a subtree which is optimal in the sense of being the smallest tree which minimizes [3]. As α increases, there arises a finite, nested sequence of optimally pruned subtrees. The estimation of how well each of these subtrees will perform if used prospectively is accomplished by cross-validation. (*See* Cart, pp 12, 75, 180, 223, 234, 306–309). Generally, tenfold cross-validation was used in Cart, but because here the end product of our calculations is an estimated survival function at each terminal node, it seems plausible that less than tenfold cross-validation will be preferred. The tree for which the first term of [3] is estimated to be smallest is an obvious candidate for prospective use. In Cart, arguments were advanced for further pruning of this initially chosen tree. In the present context a candidate for further pruning might be a further penalty for improperness of $\hat{S}^B$ if both $\hat{P}(B)$ and

$$\max_{\underset{\sim}{X}_i \in B}(C_i)$$

are large. In any case, at the termination of the computation there is a binary tree corresponding to an "optimal" partition of the range of $\underset{\sim}{X}$, and within each terminal node B a Kaplan-Meier estimate $\hat{S}^B$ of $P(T > t|\underset{\sim}{X} = \underset{\sim}{x})$, which applies to all $\underset{\sim}{x} \in B$.

The same procedures as were outlined for the Wasserstein distance functions can be applied to other metrics. One choice is the Hellinger distance between univariate distributions H_1 and H_2:

$$d_h(H_1,H_2) = \left[\int_0^\infty |\, H_1^{1/2}(s) - H_2^{1/2}(s)\,|^2 ds\right]^{1/2};$$

this metric emphasizes vertical distances, and so gives attention to the centers of distributions. We can as well employ L^p metrics, $p \geqslant 1$, on the distribution functions. With both Hellinger and L^p metrics, the horizontal truncation which applies to the Wasserstein metric and censoring becomes a vertical truncation. With the L^p metrics and proper S, δ_s is concentrated at the (or any) median for every p. Still other possible splitting criteria are these: minimal attained significance levels for the log-rank statistic (suggested by James Anderson) or values of a parametric likelihood ratio statistic least favorable to the hypothesis of equal $S^{L(B)}$ and $S^{R(B)}$ (suggested by Charles Stone).

Work on asymptotic properties of the techniques we have cited can borrow from previous efforts, since the class of sets whose probabilities figure in asymptotics here is a Vapnik-Chervonenkis class [*see* Cart, chapter 12, and Gordon and Olshen, (7)]. Very preliminary computer simulations indicate that the techniques can be useful, practical tools. We intend to report more fully in the future on detailed aspects of the algorithms and their applications.

REFERENCES

1. GOLDMAN L, WEINBERG M, WEISBERG M, ET AL. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Engl J Med 307:588–596, 1982.
2. OLSHEN RA, GILPIN E, HENNING H, ET AL. Twelve month prognosis following myocardial infarction: classification trees, logistic regression, and stepwise linear discrimination. *In* Proceedings of the Berke-

ley Conference in Honor of Jerzy Neyman and Jack Kiefer (Le Cam LM, and Olshen RA, eds). Belmont, CA, Wadsworth International Group, 1985, vol 1, pp 245–268.

3. DILLMAN RO, and KOZIOL JA. Statistical approach to immunosuppression classification using lymphocyte surface markers and functional assays. Cancer Res 43:417–421, 1983.
4. BREIMAN L, FRIEDMAN JH, OLSHEN RA, ET AL. *In* Classification and Regression Trees. Belmont, CA, Wadsworth International Group, 1984.
5. GORDON L, and OLSHEN RA. Asymptotically efficient solutions to the classification problem. Ann Stat 6:515–533, 1978.
6. GORDON L, and OLSHEN RA. Consistent nonparametric regression from recursive partitioning schemes. J Multivariate Analysis 10:611–627, 1980.
7. GORDON L, and OLSHEN RA. Almost surely consistent nonparametric regression from recursive partitioning schemes. J Multivariate Analysis 15:147–163, 1984.
8. SHILOV GE, and GUREVICH BL. Integral, Measure, and Derivative: A Unified Approach. New York, Dover, 1977.
9. BRESLOW N, and CROWLEY J. A large sample study of the life table and product limit estimates under random censorship. Ann Stat 2:437–453, 1974.
10. TSIATIS A. A nonidentifiability aspect of the problem of competing risks. Proc Natl Acad Sci USA 72:20–22, 1975.
11. AJTAI M, KOMLÓS J, and TUSNÁDY G. On optimal matchings. Combinatorica 4:259–264, 1984.

DISCUSSION

Q: (Dr. John McGuire) I don't have a question, but I'd like to do a little lobbying for these procedures on a more applied context. First, they are very useful for stratification. For example, if you have three factors, say, three levels of disease severity, two of sex, and three of age, you would have 18 separate cells in the ordinary stratification. If you apply this procedure you will very often find that you can capture all of the variability that's explained by those three factors by using, perhaps, four, five, or six cells. You may end up with one level of severity which you do not have to break down according to sex or age; in one sex, you may not have to look at age. The procedure essentially generates the minimum number of cells that you need to use to capture all of the information on the covariates that you would like to stratify by. A second use is when you're using a model you don't quite understand, which is the case 90% of the time. When you're finished and your model is fitted, the model such as the one talked about by Dr. Carter, it's useful to run this kind of tree classification on the residuals, and if you can't find any subgroup for which the residuals are significantly different from zero, then you can pass on your model-based conclusions to other people with somewhat greater comfort.

Q: (Dr. Mitchell Gail, NCI) I would like to ask a couple of questions related to the practical aspects of trying to use some of these techniques. First, are these algorithms available, and are they well-documented, so that one could know exactly what's being done? Second, you have alluded to the importance of prospective versus retrospective sampling for the training set. If this is an important topic, where can we find out about it? And third, can you give us a little bit of feeling for a comparison of how this procedure would compare with some of the more conventional procedures that have been mentioned, such as logistic regression?

A: (Dr. Olshen) The answer is yes to the first question. The software is available and documented; anyone who is interested can speak to me privately.

Retrospective versus prospective sampling in the training set is discussed somewhat in our book, from a theoretical point of view in the last chapter and a little bit from a practical point of view in Chapter 6. It's discussed in more detail in a paper that I and others wrote that's going to appear in a forthcoming Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, which will be out in August 1985.

With regard to the third question, let me say that in all the studies that we've done, comparative analyses were done with stepwise linear discrimination, with logistic regression, sometimes with projection pursuit regression, and with a variety of other things; trees don't always win. I don't have the answer to all scientific problems—none of us does—but tree methods are good a lot of the time. Generally speaking, in the kinds of studies we're talking about, they may be 10%–15% better in terms of reducing the no-data Bayes risk. Maybe they average 10% better, but they're not better every time. Logistic regression has done better in some prognostic problems in cardiology. What we're doing now is trying to combine things. In our study in making 12-month prognoses for patients who have suffered myocardial infarctions and survived to leave the hospital, we're trying to combine studies having four or five different technologies competing against three different physicians. And we're going to have the physicians vote as well as report individual prognoses and the technologies vote. The technologies voting loses information and you would think that you would be better off to try to separately estimate the probabilities of survival, average those, and plug that into a Bayes rule. We don't seem to be able to do very well at that, and just a simple process of taking votes among technologies has been useful in at least one study in cardiology.

Regression Models for Prognostic Prediction: Advantages, Problems, and Suggested Solutions[1]

Frank E. Harrell, Jr,[2,*] Kerry L. Lee,[2] David B. Matchar,[2] and Thomas A. Reichert[3,4]

Multiple regression models have wide applicability in predicting the outcome of patients with a variety of diseases. However, many researchers are using such models without validating the necessary assumptions. All too frequently, researchers also "overfit" the data by developing models using too many predictor variables and insufficient sample sizes. Models developed in this way are unlikely to stand the test of validation on a separate patient sample. Without attempting such a validation, the researcher remains unaware that overfitting has occurred. When the ratio of the number of patients suffering endpoints to the number of potential predictors is small (say < 10), data reduction methods are available that can greatly improve the performance of regression models. Regression models can make more accurate predictions than other methods such as stratification and recursive partitioning, when (*a*) model assumptions are thoroughly examined; (*b*) steps are taken (ie, choosing another model or transforming the data) when assumptions are violated; and (*c*) the method of model formulation does not result in overfitting the data. [Cancer Treat Rep 69:1071–1077, 1985]

The identification of factors related to prognosis and their use in predicting the prognosis of individual patients have importance in clinical research and clinical practice. Prognostic assessments are used in designing clinical trials, controlling for confounding factors in observational studies of treatment efficacy, counseling patients, and formulating strategies for treatment and for the optimal use of expensive medical tests.

Prognostic estimates are most accurate when based on a careful analysis of good quality data. When data have been collected rigorously, and preferably prospectively, there are many statistical techniques that can be used to derive estimates of patient outcomes. All available techniques require an adequate sample size and good patient follow-up. The sample size requirement will depend on the level of detail needed in the final model. The critical aspect of sample size is usually the number of patients suffering the endpoint of interest. When examining many potential prognostic factors, one must have a sample containing much statistical information, ie, a sample with many patients having an endpoint. Losses to follow-up should be minimal and should not be related to impending events; otherwise, prognostic estimates may be seriously biased.

Once an adequate dataset is obtained, it must be split into a "training" or model development sample, and a "test" or model validation sample. The test sample is not used at all in model development. Its purpose is to allow the investigators to assess the predictive accuracy of the derived model in an unbiased manner. Without a test sample, one cannot know whether predictions from the model will replicate on other patient samples or whether the model was "tailor-made" to fit noise or idiosyncrasies in the training sample.

After the training and test samples are defined, one selects a strategy for deriving prognostic estimates. This

[1]Supported by Public Health Service grants LM-07003 and LM-03373 (National Library of Medicine) and HL-17670 (National Heart, Lung, and Blood Institute) from the National Institutes of Health, Department of Health and Human Services, by grants HS-03834 and HS-04873 from the National Center for Health Services Research, and by grants from the Prudential Insurance Company of America, the Kaiser Family Foundation, and the Andrew W. Mellon Foundation.

[2]Division of Biometry, Department of Community and Family Medicine (*F. E. Harrell, Jr* and *K. L. Lee*), and Division of General Internal Medicine, Department of Medicine (*D. B. Matchar*), Duke University Medical Center, Durham, NC.

[3]Becton Dickinson Immunocytometry Systems, Mountain View, CA.

[4]We thank Dr. Al Bartolucci of the Southeastern Cancer Study Group and the staff of the Duke Comprehensive Cancer Center database for making the lymphoma data available, and Dr. Richard Olshen for making the recursive partitioning software available.

[*] *Reprint requests to:* Dr. Frank E. Harrell, Jr, Box 3363, Durham, NC 27710.

strategy may be regression modeling, empirical stratification on prognostic factors, or an automatic stratification scheme such as recursive partitioning (1) or entropy minimax (2). The latter two methods use a set of criteria to derive strata which separate patients into homogeneous prognostic groups that optimize their particular criteria. Each type of estimation method has its own advantages and disadvantages.

Advantages of Regression Models

Multiple regression models are being used more and more frequently in medical and epidemiological research. The Cox proportional hazards regression model (3) is often used to predict the time until a clinical event such as death, and to predict the probability that a patient will survive at least until a given time. The logistic multiple regression model (4) is an ideal method for predicting the probability of a dichotomous outcome such as hospital death or the presence of a specific disease.

Models such as these have many advantages. Perhaps most important, these models can utilize more information by relating a continuous predictor variable such as blood pressure to an outcome of interest, and by utilizing outcomes measured on a continuum, such as time until death or the severity of a disease. Stratification methods have to categorize continuous variables to be able to place patients into distinct groups. When there is a dose-response relationship between the predictor variable and the outcome, there is no way to stratify the predictor without loss of information. Regression models can also deal with multiple continuous and nominal predictors, assuming that an adequate sample size is available (or that a data reduction technique has been used, as will be discussed later).

When the regression model is formulated and its predictive accuracy is verified in the test sample, the model can be used as a concise summary of the relationships between the predictors and the outcome. Outcome estimates can be obtained for individual patients by the use of simple arithmetic in most cases. When the model does not contain many continuous predictors, nomograms can be made that allow the model to be applied easily. Pryor et al (5) demonstrated how to estimate disease probabilities from a logistic model with 9 predictors using two simple nomograms.

Problems With Regression Model Assumptions

Regression models do have disadvantages that become apparent when they are used unwisely. With the advent of powerful regression analysis computer programs, it is apparent that many researchers are treating the computer as a "black box" that can ingest data and automatically provide the right answer without the investigators using any further scientific judgment. Whereas the stratification methods can provide prognostic estimates without making many assumptions, regression methods make assumptions that must hold at least approximately for valid prognostic estimates to be obtained.

There are three types of assumptions that most regression models make: linearity, additivity, and distributional. Fortunately, the Cox and logistic models do not make strong distributional assumptions. The linearity assumption means that each continuous predictor is related to a transformation of the outcome measure in a linear fashion. The logistic model, for example, assumes that a graph of the logit of the outcome probability versus each predictor variable is a straight line, where "logit" refers to the log odds of the probability p that the event occurs, or $\log(p/(1-p))$.

The additivity assumption implies that the predictors are principally related to the outcome in an additive fashion when the model is stated in its simplest form. For example, there may be a linear effect in increasing age and an additive effect for being female which is the same at all ages. For a model containing a series of dichotomous variables, coded for example as 0 for the absence of a condition and 1 for its presence, a regression model assigns weights for each condition present and adds the weights. Thus, unions and intersections of conditions are assumed not to be the most relevant way to predict the outcome. Instead, regression models assume for example that the more symptoms a patient has, the more likely she is to have the disease.

Many regression models make an assumption regarding the statistical distribution of the response variable. For instance, the assumption that underlies ordinary multiple linear regression theory is that the response variable has a normal (Gaussian) distribution. Linear discriminant analysis goes so far as to assume that the predictor variables are jointly multivariately normally distributed. The models most commonly used for analyzing prognostic outcomes do not make such stringent assumptions. The binary logistic model makes no distributional assumptions, and the Cox proportional hazards model assumes that the distribution of failure times has the same shape for different types of patients; it does not assume a specific shape of the distribution.

When these three types of assumptions are not satisfied, the regression model may not yield accurate predicted outcomes and valid inferences about the significance of a prognostic factor. In many cases, however, the analyst can take some simple steps to satisfy the assumptions. Often transformations on the predictor variables (eg, log, square root, reciprocal) will make the relationships with the outcome linear. Additivity can be formally tested by including terms describing interaction of predictors in the model (cross-product terms). When these terms are significant, they can be retained to allow for a model that does not assume that the effect of one variable is the same at all values of the other variables. For

models requiring a strong assumption about the distribution of the response variable, transformations of this variable can often result in the assumptions being better satisfied.

Problems With Usual Regression Modeling Strategies

When a researcher is studying how a number of potential predictors relate to an outcome, one of the first problems to be addressed in regression analysis is to decide which factors to include in a model. Forward and backward variable selection are typically used to select important factors. A severe problem with these procedures is that the set of predictor variables which apparently contains the "significant" prognostic information is not unique. Harrell et al (6) show examples where the set of factors selected changes dramatically when the analysis is replicated by taking repeated samples from the same patient population. The "most significant" prognostic factor did not even remain the same in the various samples.

Even if the researcher does not care if the model is unique, the "bottom line" for a predictive method is its performance in a new dataset. Unless the sample size is large in comparison to the number of candidate predictors, stepwise regression will not yield models that predict accurately in an independent patient sample. Using the Cox model to predict time until cardiovascular death, Harrell et al (6) showed that stepwise variable selection with 30 candidate variables did not validate well when the training sample consisted of 2113 patients of whom 208 had died. Based on empirical evidence, they proposed that as a rough rule of thumb one should not attempt a stepwise regression analysis when there are fewer than ten times as many events in the training sample as there are candidate predictor variables. The problems become more severe when one considers interaction terms in the model.

Regression models (and many other methods) can fail to replicate because of the multiple comparisons problem present when one examines many candidate factors. This problem is synonymous with the multiple comparisons problems involved in comparing two treatments in multiple subgroups or in comparing multiple treatments. If one attempted to detect whether there were any differences among eight treatments by performing all $8(7)/2 = 28$ pairwise t tests, it is very likely that some of the apparently significant t tests were the result of chance (if each test were made at the alpha = 0.05 level and there were actually no treatment differences, the chance of obtaining at least one significant difference is $1-.95^{28} = 0.76$). Likewise, some of the potential predictor variables will be declared significant by chance alone. In addition, the importance of some factors will be overstated due to the overfitting of idiosyncrasies in the training sample. If one looked up 49 numbers from a table of random numbers for each of 50 patients, a regression model using the 49 random numbers as predictors would predict the times until death of the 50 patients perfectly. By definition, however, this model would have no predictive ability on any other patient sample.

TOOLS FOR STUDYING PREDICTIVE ACCURACY

Indexes of predictive accuracy are useful for judging the clinical utility of a predictive method, for measuring the amount of "overoptimism" resulting from overfitting the training sample, and for comparing the merits of competing predictive methods. If, for example, predictions in the training sample are in some sense 85% optimal but predictions in the test sample are only 70% optimal, there is evidence that the model has been tailormade to fit the training sample. If one predictive method is 80% optimal on the test sample and another is 70% optimal on the same test sample, the first method is preferred, all other qualities of the two methods being equal.

There are two major components of predictive accuracy: *reliability* and *discrimination*. *Reliability* refers to the amount of agreement between predicted and observed outcomes. If patients of a given type are predicted to have an 80% chance of surviving at least 5 years, then 80% of such patients should survive at least 5 years if the method is reliable. *Discrimination* refers to the ability of a predictive method to discern patients having good outcomes from those having poor outcomes. A method predicting an 80% likelihood of living 5 years for every type of patient may be reliable, but it would have no discrimination ability.

Reliability is an important property of a predictive method, and the reliability of any method should be validated. However, reliability is difficult to quantify because there are so many ways of subdividing patients into groups with similar predicted outcomes for the purpose of comparing the observed proportion of events with the predicted proportion of events. Discrimination, on the other hand, is easy to quantify. Harrell et al (7) proposed a general purpose index of predictive discrimination called c for "concordance." The c index is the probability that for a randomly chosen pair of patients, the predicted and observed outcomes are concordant, ie, the patient having the better outcome is the one having the better predicted outcome. When the outcome is dichotomous (eg, diseased/not diseased, death in hospital/alive), c is identical to the area under a "receiver operating characteristic" curve (8). When c is .5, the method has no discrimination ability. When c is 1.0, the method can perfectly rank patients according to the severity of their outcomes.

COMPARISON OF REGRESSION MODEL FORMULATION STRATEGIES

Introduction

Harrell et al (6) conducted a series of regression modeling and validation experiments to demonstrate the severe problems inherent with stepwise variable selection in developing prognostic models. They showed how two modeling strategies which incorporated data reduction resulted in significant improvements over the commonly used stepwise regression tactic. They began with an overall training sample of 2113 patients, of whom 208 had suffered cardiovascular death. Cox regression models (3) were developed on this training sample and on successively smaller subsets of it, to predict the time until cardiovascular death based on 30 candidate predictor variables. The models were all tested in the same test sample consisting of 2113 patients and 208 deaths. The c index was used to quantify predictive discrimination in the training and test samples. The amount of decrease in this index between training and test samples is a measure of the degree of overfitting to the training sample. The use of training samples of varying sizes afforded an opportunity to study the amount of overfitting as a function of the sample size, and to determine how large the training sample must be for a given strategy so that one can derive a model that will validate.

Stepwise Variable Selection

The first modeling method tried was standard stepwise variable selection. Variables were selected in order of statistical significance until no further variable was significant at the .05 level. The predictive discrimination in the training and test samples, for four training samples, is shown in table 1. This table demonstrates that if one wants to obtain an "apparently" good model, one should use the smallest training sample possible. However, the degree of overoptimism in this case is high; with the training sample containing only 30 deaths, c drops to .76 in the test sample. Even with the largest training sample, there is some drop in c. Stepwise variable selection will often fail to validate when the number of clinical events of interest (here deaths) is small in comparison with the number of candidate variables (say < 10:1).

Incomplete Principal Components Cox Regression

To derive accurate models from 30 candidate variables and training samples containing < 300 deaths, a data reduction technique must be applied to effectively reduce the number of candidate predictor variables. One such technique is based on computing *principal components* of the available predictors and then using a subset of these components as candidate predictors. The first *principal component* (pc) of a set of variables is that linear combination of the variables having the largest variance across patients, subject to a normalizing constraint. The second pc is the linear combination having the largest variance of all linear combinations that are uncorrelated with the first pc, and so on.

In many situations, the first few pcs will adequately summarize the variation in all the variables across patients, and these components can be used in the regression model. *Incomplete principal components regression* (9) involves fitting regression models using a subset of the entire set of pcs instead of examining all the candidate variables individually. We require the first pc to be selected first, the second pc to be added next, and so on until no more pcs are significant. The fact that pcs are examined in order of amount of variance explained (without reference to the response variable) instead of in order of p values results in a major reduction in multiple comparisons problems. In the experiments conducted by Harrell et al (6), no more than 3 pcs contributed significant prognostic information. Thus, the data reduction and the imposition of the order of selection of factors effectively substituted a small number of summary variables for the 30 individual candidate variables.

The results of the incomplete principal components regression modeling are shown in table 2. The incomplete principal components regression models validated extremely well. In fact, the model developed on a training sample containing only 30 deaths validated about as well as any model developed. This model used only the first pc and hence required the estimation of only one regression coefficient. Because of the amount of data reduction involved, the performance of this method is less dependent on the training sample size.

TABLE 1.—Predictive discrimination of stepwise Cox regression models

Training sample		Factors selected	c index	
Patients	Deaths		Training	Test
2113	208	15	.85	.82
1065	101	11	.84	.80
518	55	10	.85	.76
284	30	6	.88	.76

TABLE 2.—Predictive discrimination of incomplete principal components Cox regression models

Training sample		Factors selected	c index	
Patients	Deaths		Training	Test
2113	208	3	.84	.84
1065	101	2	.83	.83
518	55	1	.84	.84
284	30	1	.86	.84

Regression Modeling With Clinical Indexes

Incomplete principal components regression does have two disadvantages. First, these models can be difficult to interpret. Second, all original variables receive weights; it is difficult to delete variables to provide a more parsimonious model. A third technique, which solves these problems by combining features of the first two methods, was examined. The strategy has five parts. First, "variable clustering" (10) was used to group the candidate predictor variables into clusters such that within a cluster the variables were highly correlated, and variables in different clusters were as uncorrelated as possible. Second, two cardiologists made minor changes in the cluster compositions to make the clusters better reflect the pathophysiology of the disease. Third, the cardiologists assigned relative prognostic weights to the variables within each cluster. (Alternatively, the first principal component based on the group of variables can be used to derive a summary index for each cluster.) Fourth, minor adjustments were made in weights found to be significantly inadequate in the training sample by a Cox regression analysis. The weighted variables in each cluster formed a clinical index. This process reduced the 30 original variables to 5 clinical indexes and 5 variables which stood alone. Fifth, stepwise variable selection on the resulting 10 candidate factors was used to derive the final model for each training sample.

The results of the variable clustering/clinical index strategy are shown in table 3. Except for the smallest training sample, the regression models using indexes performed almost as well as the incomplete principal component models. The cardiologists found these models to be readily interpretable, because the interpretation was built into the model design.

Comparative Summary of Regression Strategies

Even using a training sample consisting of 2113 patients of whom 208 had died, stepwise variable selection using 30 candidate variables (still $> 1/10$ of the number of events) had serious problems with predictive accuracy and interpretability. Incomplete principal components regression and regression on indexes, which resulted in a data reduction to 1-3 and 10 candidate variables, respectively, in these examples, performed better. Regression on principal components had the best predictive discrimination in the test sample. Regression on cluster indexes was almost as good and the models were easier to interpret.

TABLE 3.—Predictive discrimination of Cox regression models using clinical indexes

Training sample		Factors selected	c index	
Patients	Deaths		Training	Test
2113	208	8	.85	.83
1065	101	6	.83	.83
518	55	5	.85	.82
284	30	3	.86	.80

COMPARISON OF REGRESSION MODELING WITH OTHER METHODS

As the experiments discussed above demonstrate, the actual regression modeling strategy can have a significant effect on the predictive accuracy of the final model. It would be of further interest to compare the accuracy of different regression modeling strategies with methods that are much different from regression. To this end, experiments were conducted to compare three regression methods with the method of recursive partitioning (1). Unlike the above examples, the endpoint selected is binary instead of continuous. Hence, the logistic regression model (4) was the regression model chosen to predict the probability of occurrence for this endpoint.

Dataset

The data for these analyses were obtained from the Southeastern Cancer Study Group. Twenty-nine institutions participated in protocols for the treatment of patients with non-Hodgkin's lymphoma. The training sample consisted of 334 patients who participated in the cyclophosphamide, vincristine, and prednisone (COP) protocol versus COP plus carmustine (BCOP) protocol from 1973 to 1977. The test sample was derived from data on 296 patients followed from 1977 to 1982 on the BCOP versus COP plus doxorubicin (CHOP) protocol. Patients in the training sample receiving either treatment were grouped together, as specific treatment appeared to be less important than other prognostic variables (11). On the other hand, only patients in the BCOP arm of the BCOP/CHOP protocol were included in the test sample, as CHOP therapy appeared to be significantly better than BCOP in certain subgroups (11). Information was organized and coded by nurse coordinators using explicit definitions for categorical variables. Complete response (CR), defined by absence of disease at the end of six cycles of chemotherapy, and subsequent survival were assessed. The endpoint used here is the initial response, CR.

Twenty-five candidate variables were considered. Seven of these were continuous, one was ordinal, and the rest were dichotomous. The overall training sample was divided into two mutually exclusive training samples. The first consisted of 110 patients, of whom 50 had CR. The second consisted of 224 patients, of whom 102 had CR. The test sample in all cases had 116 patients, of whom 38 had CR.

Regression Methods

Three regression modeling strategies were used to develop a logistic model for predicting the probability of CR for each training sample. First, standard stepwise variable selection using the SAS LOGIST procedure (12) was performed. Second, a logistic regression model using a single prespecified index was derived. This index, which we will call a "sickness score," was derived without reference to the data and with no input from a clinician. It is defined by simply counting the number of conditions present from among the following: prior treatment, stage IV, loss of > 10% body weight, fever, night sweats, other symptoms, bone marrow involvement, node involvement, CNS involvement, spleen involvement, abdominal involvement, diabetes, obesity, and unfavorable histology. It is important to note that no attempt was made to add any of the important continuous variables into the sickness score; only the dichotomous variables were considered. The third modeling method was incomplete principal components regression (see above).

Recursive Partitioning

The method of recursive partitioning (1) was used on the same dataset to derive estimates of the probability of CR. This is done by sequentially examining all possible stratifications to discover the variable which, at some value, best predicts outcome. As this examination is repeated (recursively), a tree is grown until a defined stopping rule is met. The resulting "bushy" tree is "pruned" by a cross-validation procedure to compensate for overfitting. For patients who fall into a particular terminal node of the tree, the probability of CR is estimated from the experience of members of the training set (simple resubstitution proportions). The branches of the tree are defined by intersections of conditions. All continuous variables must be classified into a small number of groups, for example, age < 40, age 40–53, age > 53.

For each predictive model and recursive partitioning decision rule, the predictions were evaluated in the test sample by calculating the c index to measure predictive discrimination.

Results

The results of applying each predictive method to the test sample are shown in table 4. Stepwise regression performed poorly using the smaller training sample, but performed better when based on the larger training sample. The logistic model using only the sickness score and based on either training sample performed relatively well. The incomplete principal components logistic model derived from either training sample was superior. Recursive partitioning based on the smaller training sample had better discrimination than stepwise variable selection, but was not as good as the models using principal components or the sickness score. Recursive partitioning using the larger training sample surprisingly performed very poorly. An alternative stratification method, entropy minimax (2), was also used. Entropy minimax based on the larger training sample performed much better than recursive partitioning, and as well as the best regression procedure (incomplete principal components logistic regression). A detailed comparison of regression and stratification methods applied to these data is in preparation.

Cook and Goldman (13) compared recursive partitioning and logistic regression in predicting acute myocardial infarction. They found that recursive partitioning had a higher c index than did the usual stepwise logistic regression strategy, in agreement with our findings for the smaller training sample.

CONCLUSION

Multiple regression models have wide applicability in problems involving the assessment of patients' prognoses. They can be used to predict dichotomous, polychotomous, ordinal, and continuous outcomes using any mixture of dichotomous, polychotomous, and continuous predictor variables. Models such as the Cox proportional hazards model can handle "censored" responses occurring when one is interested in the time until the occurrence of an event and not all patients have suffered the event at the time the data are analyzed.

Regression analysis does have definite limitations. The model chosen must "fit" the data. Models frequently fail to validate due to overfitting the training sample when the number of potential predictors is large in comparison to the number of events. Regression analyses must be performed with care; the analyst must resist the temptation to throw the data into a computer program without constructing graphs and performing formal statistical tests to check the model assumptions. Furthermore, she cannot ignore the problems of multiple comparisons.

It is important to employ a data reduction technique before doing a regression analysis when the number of events in the training sample is smaller than roughly ten times the number of potential predictor variables. The data reduction should be done with no reference to the re-

TABLE 4.—Predictive discrimination of 3 regression methods and recursive partitioning in the test sample

Method	Training sample used	
	n = 110	n = 224
Stepwise variable selection	.58	.67
Sickness score	.66	.64
Incomplete principal components	.67	.68
Recursive partitioning	.61	.56

sponse data. In particular, the commonly used method of "univariable screening" is not appropriate because it fails to take multiple comparisons problems into account. Some methods of data reduction are (*a*) deleting variables using prior knowledge of the problem, (*b*) deleting variables known to have high measurement error, (*c*) principal components analysis, and (*d*) deriving summary indexes after clustering variables either on statistical or clinical grounds. As shown above, data reduction can have an important impact on the quality of the predictions.

When a regression analysis is not performed carefully, other methods such as recursive partitioning can be more accurate. When the analysis is done carefully, however, regression models can make prognostic estimates that are often as good as and sometimes better than those obtained using stratification methods.

REFERENCES

1. BREIMAN L, FRIEDMAN JH, OLSHEN RA, ET AL. Classification and Regression Trees. Belmont, CA, Wadsworth International Group, 1984.
2. CHRISTENSEN R. Entropy Minimax Sourcebook, Volume I: General Description. Lincoln, MA, Entropy Limited, 1981.
3. COX DR. Regression models and life-tables (with discussion). J Royal Stat Soc B 34:187-220, 1972.
4. WALKER SH, and DUNCAN DB. Estimation of the probability of an event as a function of several independent variables. Biometrika 34:167-179, 1967.
5. PRYOR DB, HARRELL FE, LEE KL, ET AL. Estimating the likelihood of significant coronary artery disease. Am J Med 74:771-780, 1983.
6. HARRELL FE, LEE KL, CALIFF RM, ET AL. Regression modelling strategies for improved prognostic prediction. Stat Med 3:143-152, 1984.
7. HARRELL FE, CALIFF RM, PRYOR DB, ET AL. Evaluating the yield of medical tests. J Am Med Assoc 247:2543-2546, 1982.
8. HANLEY JA, and MCNEIL BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 50:23-36, 1982.
9. JOLLIFFE IT. Discarding variables in a principal component analysis. II. Real data. Appl Stat 22:21-31, 1973.
10. RAY AA, ed. SAS User's Guide: Statistics. Cary, NC, SAS Institute, 1982, pp 461-473.
11. OLSEN GA, REICHERT TA, and BARTOLUCCI A. Specific effect of Adriamycin on survival in diffuse aggressive non-Hodgkin's lymphoma. Blood 64(suppl 1):183a, 1984.
12. HARRELL FE. The LOGIST procedure. *In* SUGI Supplemental Library User's Guide, 1983 ed (Joyner SP, ed). Cary, NC, SAS Institute.
13. COOK EF, and GOLDMAN L. Empiric comparison of multivariate analytic techniques: advantages and disadvantages of recursive partitioning analysis. J Chronic Dis 37:721-731, 1984.

PANEL DISCUSSION: Biometric Issues

Panelists: Drs. John Crowley, Edmund Gehan, and Larry Norton

Dr. John Crowley: I would like to thank all the speakers for taking time to try to help educate us and to thank the organizers for putting together the conference. I'm going to be a little bit argumentative, in part because I think that's what I'm being paid to do, and in part because of a little psychology that I want you to understand. There was an occasion several years ago, a similar occasion, in which some relative outsiders (and excuse me if that definition doesn't fit all of you) were brought in to talk to some of the in-club people about some new ideas in statistics and how they might be used in clinical trials. I said a few things about using my test for ordered alternatives, Ross Prentice said a few things about relative risk regression, Butch Tsiatis said a few things about group sequential methods, and Richard Peto, who is the insider, got up and said all they have to do is randomize, stratify retrospectively, and do log rank tests. That was his discussion of our papers. And that's kind of the psychology that I'm tempted to adopt. I'm explaining this so that you'll modulate my remarks a little bit.

Now that I feel more like an insider I have to ask: Do I need all this in doing cancer clinical trials? I wonder. Regarding the first speaker, Dr. Coldman, my first thought, in spite of the fact that I like modeling, was: Is that all there is? Models, of course, are never realistic enough; we can always pick them apart. The trick is to get one that's both tractable and, at the same time, rich enough so that you can learn something from how the model fits and how the model doesn't fit. I'll challenge Dr. Coldman to contradict me. I didn't get that sense from the presentation. If combination chemotherapy has agents that act differently, the thing to do is to use them together. If they can't be used together because of toxicity, they have to be alternated. I didn't see that the model introduced new principles or elucidated old ones in that regard.

Dr. Carter and I have had this discussion at another forum. The response surface designs that he introduced are factorial but more complicated in the analysis than the ones that Dr. Byar mentioned. Some of these points came out in the discussion. But for reiteration, they're hard to do, and they're hard to analyze. Assumptions are made and dangers are introduced, in particular, the danger of model misspecification. More importantly, I'm not convinced of the relevance to cancer clinical trials, and I'm waiting for the evidence from the ongoing trial to contradict me. It seems to me that many of the cells in the design Dr. Carter proposed could be ruled out on a priori grounds; others could be ruled out with fewer patients by early phase I and phase II studies. And we should save the big randomized studies for more comparative rather than optimality questions. I feel especially strongly about that if we're talking about different regimens. The example given is one of dose and schedule, which I think is a little bit more interesting and promising, but I would suggest that many of the cells could have been ruled out.

The discussion on factorial designs I found interesting. I have to tell another little anecdote. As discussants, we did get various amounts of information from each speaker prior to coming here. This one was an exception. It reminds me of what Ross Prentice said when he had a similar situation, when Paul Meier failed to come through with an advanced copy of the remarks. Ross said something to the effect of: "Factorial designs, that's a good title." I was quite struck when I first read Peto's and then Norm Breslow's arguments about the fact that we should be thinking more about the opportunities for 2×2 and $2 \times 2 \times 2$ designs. Having been trained in, and having learned a lot about, the agricultural designs, and then having forgotten them for several years in clinical trials work, I was rather embarrassed. I have been trying ever since to find good ways to use 2×2 designs and have yet to find one.

The trouble seems to be that either non-A or non-B is prohibited or AB is prohibited. Dr. Byar suggests the possibility of A,B versus AB, but now it's pretty clear that we're not getting twice the advantage for the same number of patients. So I'll continue to be diligent, and, in particular, when someone comes up with an innocuous add-on study (like manicure styles for different kinds of cancer patients) that I'm sure won't interfere with the major questions that are being asked, I will think of a 2×2 design.

I was, in fact, prepared to be the most argumentative about recursive partitioning, mainly because I don't understand it very well. In fact, let me shift gears a little bit. I'm really quite excited about the whole area as a topic for statistical research, but let me back off a minute to say that I was glad to hear some discussion from both these speakers about how hard it is to do, how hard even regression is to do, much less something which seems to be even fancier, like recursive partitioning. And I'll repeat the comment that this subject has relatively little do do with covariate adjustment in clinical trials of the kind that we're used to doing, where the adjustment is relatively unimportant, due to the randomization. Regression is hard enough without censoring, without binary outcomes, where you can't get a good picture. It's almost impossible to make any sense out of a single observational study and individual factors that show up in that study. I think those points came out quite clearly from those speakers. But I am quite interested in the idea of putting people into bins for stratification, as one questioner mentioned, for staging criteria. That's exactly the kind of thing, it seems to me, that we want to do and that's why I was attracted to the recursive partitioning algorithm. We want to put people into not just two bins but several bins with an ordered kind of classification scheme, 1,2,3,4, and usually our variables are the more complicated ones that Richard Olshen alluded to at the end of his talk, that is, survival in which there is censoring. So there's a lot of work to do, but I do think that's an interesting area and I will just close by saying, there's validation and validation. I have, in fact, had some experience in doing both Cox regression and some variation of recursive partitioning to put people into bins and was fortunate enough to have an independent test sample from another institution; both methods did very well on the original data, on reuse data; they did absolutely horribly on the other institution's data. Dr. Olshen had an example where that was not the case. I think this is something to keep in mind: no matter how many times you reuse the same sample, it's still the same sample.

Dr. Edmund Gehan: First, I'd like to thank Rich Simon and Susan Ellenberg for organizing the session on Biometric Issues and inviting me to be a discussant. Following John Crowley's lead, I'll try to be a little bit argumentative in discussing the papers also.

I've read and heard a lot about the Goldie-Coldman hypothesis, but this is the first time I've heard it presented. *My* hypothesis is: "The Goldie-Coldman hypothesis is really the same as the Skipper-Schabel hypothesis." What Goldie and Coldman have been doing in logical and mathematical terms, Skipper and Schabel have been doing in experiments in animals. In nearly all of the experiments by Skipper and Schabel, they postulate that a series of administrations of treatment lead to constant percent killing of tumor cells, the constant depending upon the dosage. Their work is most relevant when tumors are fast-growing, and its best application has been found in clinical studies of lymphoma and leukemia. Consequently, I believe that the Goldie-Coldman hypothesis will have most relevance in lymphoma and leukemia studies.

In a paper of Dave Byar's, he categorizes prognostic factors as follows: tumor factors, host factors, and those factors resulting from the effect of the tumor on the host. In Dr. Coldman's talk, he considered only characteristics related to tumors. He did indicate that he was taking account only of factors associated with tumors; however, I wish to emphasize the fact that the other types of prognostic factors can be of great importance, especially in the development of regression models.

One of the features that I find unsatisfactory about the Goldie-Coldman hypothesis is that it doesn't permit detailed testing. Basically, the Goldie-Coldman hypothesis has resulted in rather broad recommendations for strategies of treatment; eg, administering sequential non-cross-resistant regimens of treatment would be expected to yield better results than, say, continuing a single combination treatment regimen. The only test of their proposed strategy would be the general one of comparing the outcomes of alternative

treatment programs. Since their model was developed from certain logical and mathematical principles, one might have expected that it would be subject to a more specific test. One advantage of a mathematical model is that it should provide a prediction outside the range of experimentation. Also, when a mathematical model is fit to some actual data, there is always an "error term" which should follow a certain statistical distribution, depending upon the specific model. The Goldie-Coldman model doesn't lend itself to this specific kind of test. Having said that, it should not be too difficult to test their broad recommendations for treatment, such as administering sequential non-cross-resistant regimens for patients with various types of cancer. In one recent study by Bonadonna et al in Hodgkin's disease, giving sequential MOPP and ABVD led to better results than either combination treatment administered alone, with respect to remission rate, length of remission, and survival. In that circumstance, the Goldie-Coldman hypothesis was broadly confirmed, but I'm concerned that the model is not amenable to more detailed testing.

In Dr. Carter's paper, he has introduced response-surface type designs to clinical studies that George Box and co-workers introduced into industrial experiments in the 1950s. Perhaps one of the reasons that Box's ideas did not find application in clinical studies earlier is that there are important differences between experiments in industry and in the clinic. In many industrial testing situations, there generally are a larger number of treatments, the experimental units are usually homogeneous, and there is a capability of repeating experiments within relatively short periods of time. If the first experiment in a sequence provides one set of results, a later experiment could be designed to seek the optimum response surface. However, in clinical trials the experimental units (patients) are heterogeneous, and it ordinarily takes a fairly long time to observe an outcome for each patient. Consequently, one would expect Box's ideas to be less applicable in clinical trials than in industrial experimentation.

One question raised in the discussion of Dr. Carter's paper was: Is this a phase II½ study in the context of phase I, phase II, and phase III trials? I think that the response-surface type design may be applicable in phase IV clinical studies. If a phase III trial has established a regimen as effective, one would ordinarily want to find out the optimum way of administering the regimen. Does anyone really believe that the MOPP regimen that was originally proposed in Hodgkin's disease has the optimum dosage and schedule for the various agents? I suspect that when a treatment regimen has been demonstrated as beneficial, it tends to become accepted as it was proposed and there is not much investigation of the optimum dosages, routes, and schedules. I believe that more of this should be done. This is especially true when there are no very new ideas for treatment regimens to test in phase III trials. Such patients could contribute more to clinical knowledge by being entered into a phase IV study which was designed to determine an optimum way of administering a regimen. There is one practical problem that arises even in such a simple clinical trial as a comparison of a high versus a low dosage of a treatment in a particular type of cancer. In some clinical trials that I have worked on, patients receiving the higher dose are taken off therapy much earlier than patients who are receiving the low dose. Hence, the high dose becomes an intermediate dose. This tends to occur when there is too large a difference between the high and the low dose. Finally, the question of covariates was also raised in Dr. Carter's talk. With the multiplicity of variables in Cox's model in that circumstance, I think it would probably be better to stratify patients with a random assignment of patients within each stratum rather than to use a modeling approach with a covariate for each prognostic feature. I'd be interested in hearing Dr. Carter's comments.

Dr. Byar has challenged us to come up with an example of a factorial design, and I recall a clinical trial that was performed in the late 1950s and published in 1961.[1] This study was performed by a cooperative group that was then called the Acute Leukemia Group B, and the design was "partial factorial," the two factors being 6-MP (yes or no) and methotrexate treatment (yes or no). The study was partial factorial because the "no treatment" regimen

[1]Frei E, et al. Studies of sequential and combination antimetabolite therapy in accute leukemia; 6-MP and methotrexate. Blood 18:431-454, 1961.

was not included and the comparison was among three treatments: 6-MP versus methotrexate versus 6-MP plus methotrexate. The treatments were compared for their capability in inducing remissions in acute leukemia. Patients were not maintained in complete remission, so that when they relapsed, the 6-MP patients received methotrexate and the methotrexate patients received 6-MP. This was a very interesting study because the response rate to the combination was very close to that predicted assuming independent action of 6-MP and methotrexate. This was an early study that suggested that combination treatment resulted in a higher remission rate than either agent administered individually, and that led to bigger combination regimens being tested in later studies Further, response rates to 6-MP and methotrexate in phase II were nearly the same as those in phase I, so that response in the second phase was nearly independent of that in the first. This is an early example of a partial factorial design in clinical cancer research.

I found Dr. Olshen's paper very interesting. I believe that the recursive partitioning approach will handle very well situations that cause problems when a regression modeling approach is taken. An example is the problem of the very small subset of patients; eg, there are 200 patients in the study and five or ten have some condition, say, a very poor performance status or abnormal value for a particular test. If the frequency of such an event is very low in the whole population, a regression modeling approach doesn't really work very well, but the recursive partitioning approach would seem to handle that situation perfectly. In classification, I would distinguish between approaches to classifying diagnoses and prognoses and wonder if recursive partitioning isn't better suited to problems in diagnosis rather than prognosis. For example, in diagnosis patients might be classified as having had a myocardial infarction or no infarction, whereas in prognosis we might be interested in predicting survival or death within some period of time. With these tree-structured methods, you end up with a classification, eg, being a long survivor or having had a myocardial infarction. For problems in predicting prognosis, the logic of a regression approach seems superior to me to the recursive partition approach. The outcome of a logistic regression analysis would be a prediction of the probability of surviving 2 years, say, whereas the tree-structure method would result in a statement that "you are going to survive 2 years"; however, there is a certain chance of error in making that statement. Finally, I'd be interested in knowing something about the computing cost of the recursive partitioning approach versus a regression modeling approach.

One question for Dr. Olshen is: How do you test for a treatment effect in a recursive partitioning model? In a regression modeling approach, a treatment term could be added and a statistical test could tell us whether adding a term for treatment resulted in a significantly better fit to the data. It is not very clear to me how this would be done in a recursive partitioning situation.

This is the second opportunity that I have had to discuss Dr. Harrell's paper, since I was a discussant at a recent meeting of the American Statistical Association. In summary, I found this paper to be a very useful piece of work. I have a question concerning the possibility of testing differences among C indices. In the example relating to heart disease, one approach resulted in an index of .84, whereas another yielded .76. Certainly, .84 is higher than .76 and possibly should be preferred for that reason, but it would be useful to have some statistic for testing the difference (or the ratio) of C indices. With all these models, and Dr. Olshen has emphasized this, evaluation of a particular modeling approach comes down to the use that we make of it.

In relation to Dr. Harrell's talk, clinicians have often asked me whether it is worthwhile doing one more laboratory test, say, an immunological test or a measure of the pharmacological action of a drug. If you were to attend a medical meeting, you might hear a research paper on: Is the result of a particular immunological test predictive of long survival? or, Is it predictive of response? In a regression modeling approach, one can put a term in the model for the result of a proposed test and evaluate whether or not addition of the term results in a better fit to the data. In the approach based upon incomplete principal components of variable clusters, any new immunological test will tend to get lost in one of the components or clusters. Yet, the new test may have a status such that it should be considered by itself. I think that a common problem in clinical applications is: Can one more

variable be added that will help us in making predictions? Finally, I'm interested in the computing cost of Dr. Harrell's work. He has done a tremendous amount of computing. It seemed to me that the incomplete principal components approach probably costs the most, followed by variable clustering and then selection. I'd be interested in Dr. Harrell's relative cost evaluations.

Dr. Larry Norton: I would like to add my word of thanks to Dr. Simon and the other organizers of this symposium for the opportunity to be here and join the discussion. It has not escaped my notice that I am the nonstatistician on the podium. In addition, I think, perhaps for the first time in my career, I will be the least argumentative person on the podium, although I do not think that these two facts are related. I would like to use my time to stress the common thread that I believe runs through all of the papers presented here today. This theme is one that is highly relevant to those of us involved in therapeutic trials from the clinical, as opposed to the analytical, end. It concerns a changing notion of the meaning of a clinical trial.

A trial is often thought of as a contest, pitting one type of treatment versus another, or others, with the hope that one therapy will emerge as victorious to replace all others in clinical use. Certainly this concept has advantages, not the least of which is the ease of extrapolation to actual practice. We may take the "best" arm, however that is defined, and immediately use it to treat our patients. But can we be sure that the experimental results apply to the particular patient that we are treating? Also, unfortunately, victors are not always leaders. That is, having found that a certain therapy results in a superior remission rate or duration, or some other variable of clinical importance, we are left with the question: Where do we go from here? The answer is usually another contest, between the now "standard" therapy and another approach that may not build on the lessons of the previous trial.

Laboratory science, in contrast, is seldom conducted as a series of contests. Large, multifactorial designs are often used, results are examined minutely, and new hypotheses are generated by inspection, analysis, and repetition of crucial items. Subtle variations in experimental technique may be used in the design of new experiments, conducted swiftly and efficiently with rapid turnaround of data. The experiment, therefore, is an aid to reasoning, not just the definitive culmination of reasoning. This approach is, of course, difficult or impossible in the clinical setting, because of not only ethical constraints but also those of time and money. Yet the spirit of the laboratory should be kept in mind as we examine the modern clinical trial.

What we have heard today are ideas that do reconsider the clinical trial from this point of view. Dr. Coldman's paper has been extensively discussed. I myself have problems with some of the asssumptions of the model, and hence its clinical implications. Yet I defend the use of biomathematical models for generating therapeutic ideas. It is reasonable here, as in any area of science, to start with simple systems and basic assumptions. It is important, however, not to be constrained by these assumptions. This comment refers specifically to the "error" term mentioned by Dr. Gehan. If, in clinical use, the model fails to predict real events, we must carefully examine the correspondence between these assumptions and our actual situation. I am particularly concerned with the requirements of equal growth rates of equally sized subpopulations of cells. I believe that "alternating non-cross-resistant" chemotherapies will work on occasion, perhaps in those cases where these stringent criteria are applicable. If, however, this approach does not work in a given situation, is it because of a significant deviation from these criteria? This question can be addressed only by a careful analysis of a clinical data-set with regard to determinants of response. Groups of pretherapy prognostic variables may disclose underlying biological states relevant to therapeutic outcome. A restricted view of the clinical trial as a contest between the modeling idea and a standard control may obscure these important observations.

On the other hand, we must not be so flexible in our analysis that we mold our hypothesis to fit our conclusions. In the last several years we have seen many clinical programs, not all of them real trials, that have been labeled with the term "Goldie-Coldman" although they actually apply already well-established principles of combination chemotherapy. This is unfair to the model. Authors of a treatment program must be very specific in their use of

ideas so as not to attribute success or failure indiscriminately.

A well-designed clinical trial should be interpretable from the viewpoint of the model: The answer may not be "yes" or "no," but a carefully reasoned "sometimes." A difficulty with this approach is its inefficiency, the need for large data-sets or multiply repeated experiments. For example, the alternating use of two combinations, MOPP and ABVD, has been reported to increase the response rate and duration of response in advanced Hodgkin's disease. If, on replication in another setting, such superiority is lacking, perhaps the patient populations differ in terms of biological adherence to the assumptions of the model. To be tested this would require another trial with a more carefully selected patient group. This process is not only slow and expensive, but also scientifically dangerous: the results of the previous trial may influence referral and accrual patterns to the successor study. Not all prognostic factors can be specified in an eligibility list, for subtle determinants may exist. These include compliance and other psychological factors. If a positive trial now attracts patients of poorer prognosis, good results may deteriorate. Alternatively, if patients with certain identifying variables seem not to do well on MOPP/ABVD, referring clinicians may exclude these patients from future studies of this regimen, so results may improve.

As we move away from large, simple randomization plans into more complicated schemes, interactions between selection and prognosis may get more important. Indeed, interactions between variables may get more important as we increase the number of these variables. For example, the response-surface design described by Dr. Carter seems to presuppose a great deal of knowledge about the regimens. However, important facts may be unknown prior to conducting the trial. At doses of methotrexate that do not require leucovorin, there is no possibility of an interaction between 5-FU and leucovorin. At intermediate doses of methotrexate, the leucovorin may increase the toxicity of 5-FU and thereby hamper its clinical efficacy. As the dose of methotrexate increases, the benefits or toxic effects of the higher dose may obscure any leucovorin-5-FU interactions. Variations in time between doses may create further complications. The important point is that these interactions may not be apparent in the marginal distributions, so that the sizing of such a factorial design may be quite difficult before the fact. Also, if the marginal distribution is multimodal, there is no a priori reason to suppose that we will pick up the optimal peak. The same applies to a local minimum versus the global minimum of therapeutic benefit.

Recursive partitioning is potentially subject to similar dangers involving interactions. As this statistical research evolves we must pay particular attention to the influence of therapy on prognostic variables. Today's speakers have impressed me with the difficulty of these analyses. I am concerned as well with the difficulty of biological interpretation should these analyses prove feasible. Ultimately we may be left with the issue of "validation." As presented by Drs. Olshen and Harrell, validation implies repeated trials, which means that patient selection factors may come into play. Repeated trials may also be hampered by flagging investigator interest ("Why repeat a study that's already successful?") and poor accrual because of ethical questions ("How can this patient be randomized to a treatment 'proven' inferior in a previous study?"). Such is the price we pay as we go beyond the clinical trial as contest.

Regarding Dr. Byar's presentation of factorial designs, I am currently involved with a very large, prospectively randomized trial of postoperative adjuvant chemotherapy of breast cancer (CALGB 8082). This trial fits a factorial design in that two different versions of initial chemotherapy (CMFVP) are used, with a second randomization at 8 months to continuing the CMFVP or switching to a doxorubicin-based regimen. So far the study is progressing nicely, with an accrual in excess of 900 patients. It will be analyzed for interactions between the CMFVP and the cross-over regimen, with the possibility of collapsing groups for pairwise comparisons. This is, of course, a simpler design than those we have heard presented today, but it has proven feasible in practice.

My last comment concerns Dr. Harrell's "sickness score." This is a tremendous victory for clinical intuition: You just combine all of the negative prognostic impressions and come out with something highly predictive. My only suggestion is not to turn to the physician for this judgment because, as we have known for a long time, the nurses who actually

care for the patient are really much better at it. We should recognize that prognostic variables, including subjective ones on the part of the patient and the health care team, are highly interrelated: As we compartmentalize our experience we must never lose sight of the whole picture. I like the "sickness score" best of all.

Size of Phase III Cancer Clinical Trials

Richard Simon[1,*]

An effective clinical trial must ask an important question and provide a reliable answer. A major determinant of the reliability of the answer is the size of the trial. Trials of inadequate size may cause contradictory and erroneous results and thereby lead to the inappropriate treatment of patients. There is no single answer to the question of how large a phase III trial should be. But there are important factors to be considered in establishing the size and duration of a trial, and it is my purpose here to review them.

Statistical Basis for Sample Size Calculations

I will begin by summarizing the statistical framework commonly used for sample size calculations in a randomized study comparing a new treatment to a control. To simplify matters, first assume that there are no patient subsets of interest for separate analysis. Sample size calculations are generally based on three parameters: α, power, and Δ. The first, α, denotes the type 1 error. It is assumed that at the end of the trial an α level statistical significance test will be performed to test the null hypothesis that the treatments are of equal efficacy. The type 1 error represents the probability that the two treatments will be erroneously determined to be different in efficacy if in fact they are equivalent. A value of .05 for α is generally used. The power represents the probability of obtaining a statistically significant difference at the end of the trial if the new treatment is in fact better than control by an amount Δ. Clearly, to determine sample size, one must define the endpoint and scale of measurement and then select values of power and Δ. For example, survival or disease-free survival could be the endpoint in adjuvant studies, and the scale of Δ could be the difference in 5-year survival proportions or ratio of median survivals. The technical aspects of scale of measurements are not of fundamental importance to the main issues here. The main point is that for a given endpoint, scale of measurement, and power for detecting a difference of size Δ, the required number of patients depends critically on Δ. This can be seen in table 1, where the number of patients per treatment for detecting a difference in 5-year survival rate is tabulated as a function of Δ and the expected 5-year survival rate in the control group (1). In this table, a two-sided α of .05 and a power of .90 are used. It is assumed that analysis occurs after all patients are followed for 5 years.

How Should the Difference To Be Detected (Δ) Be Chosen?

The parameter Δ does not represent the difference we hope exists; it represents the size of difference that the study will be able to detect. For major phase III studies, it should generally represent the smallest difference that would be of clinical significance. For some pilot studies, it may be reasonable to have a larger Δ. Many "negative" studies are misinterpreted due to the use of significance tests rather than confidence intervals and the use of unrealistically large Δ's. Lack of a statistically significant difference does not demonstrate that the treatments are equivalent; it only demonstrates that we can reliably rule out differences of Δ or greater. If Δ is chosen very large, then nonstatistically significant results will be consistent with both no difference and with major differences that are less than a "home run." For example, a complete response rate of 25/50 (50%) is not significantly different from one of 20/50 (40%) ($P = .42$), but an approximate 95% confidence interval for the true difference ranges from 13% favoring one treatment to 33% favoring the other treatment!

Although there have been a few home runs in cancer therapeutics, they are often apparent regardless of the methodology employed. Randomized clinical trials are needed to identify important advances that are not home runs and to develop stepwise components of major breakthroughs. An absolute 15% increase in survival for a common type of cancer represents a major public health accomplishment that has rarely been achieved in oncology (2). For example, with about 100,000 new cases of stage II breast cancer in the United States, United Kingdom, and Western Europe per year, an absolute increase

[1]Biometric Research Branch, Division of Cancer Treatment, National Cancer Institute, Bethesda, MD.

[*]*Reprint requests to:* Richard Simon, PhD, Biometric Research Branch, Division of Cancer Treatment, National Cancer Institute, Landow Bldg, Rm 4B06, Bethesda, MD 20205.

TABLE 1

Survival Proportion In Control Group	Difference Δ In Survival Proportions .10	.15	.20
.05	198	106	69
.10	281	143	89
.20	410	197	117
.30	496	231	133
.40	533	244	141
.50	533	242	133
.60	496	216	117
.70	410	172	89
.80	281	106	-

Number of patients per treatment to have power .90 for detecting difference Δ in survival proportions. Type 1 error of α = .05 (2 sided).

of 15% in 5-year survival benefits 15,000 new patients every year. If the 5-year survival rate for such patients on the control treatment is about 60%, then a 15% absolute increase represents a benefit for 37.5% of the patients who could have benefited. Based on current concepts of the probability that a solid tumor has developed multiple resistance by the time of diagnosis, and based on the difficulty of substantially extending survival compared to temporarily shrinking a solid tumor, it may not often be plausible to expect survival differences of > 15% for adjuvant therapy studies of current cytotoxic therapy. Overall results for the adjuvant chemotherapy of premenopausal stage II breast cancer do not represent a > 15% absolute increase in 5- or 10-year survival proportion.

Differences in survival proportion of < 10% are very difficult to seek in a single study. Therefore, it seems important to design at least some major trials to have 90% power for detecting differences in survival proportion of 10%–15%. If disease-free survival is the primary endpoint, a similar argument for using these parameters can be made, although differences of 20% have been seen in primary breast cancer. Using table 1, this generally implies a sample size in the range of 150–500 patients per treatment group for each set of patients that will be separately analyzed.

For diseases in which survival is very short, it is not natural to speak in terms of differences in the proportion of patients who survive beyond a specified time, and the above guidelines may not apply. In general, for studies with survival or disease-free survival as endpoint, unless the curves show a clear plateau, then planning the trial based on comparing the entire curves is a better approach. Table 2 shows the approximate total number of deaths necessary to provide power .90 for detecting a specified ratio of median survivals (3,4) for exponential distributions. For example, to detect a 50% increase in median survival, it is required that the combination of number of patients accrued and duration of follow-up must be such that a total of about 256 deaths are observed by the time of analysis. The 50% increase in median survival corresponds to at most a 15% absolute difference in proportion surviving beyond a specified time (eg, 5-year survival proportion). Similarly, the 30% (1.3) and 75% (1.75) increases in medians correspond to maximum absolute differences in survival proportion of about 10% and 20%, respectively. To observe 256 deaths in a disease with a long median survival, many more than 256 patients need to be entered if the study is to be completed in a reasonable period of time. For diseases with very poor prognosis where median survival is < 1 year, it may be reasonable to target the smallest ratio of medians of clinical significance to be ≥ 75%. For example, if the median survival is only 6 months, it may be reasonable to target a doubling of median survival to be the Δ of smallest clinical significance. Since it would be feasible to follow all patients until failure in this circumstance, a clinical trial of only 88 total patients might be appropriate

(table 2). But for other types of cancer, a 50% increase in median represents a very major effect, and larger differences may not be plausible. For example, newly diagnosed prostatic cancer patients with stage D_2 disease have a median survival of about 3 years. A 50% increase represents an increase of 1.5 years in median and 2.2 years in mean survival. To have 90% power for detecting an increase of this size, a study with about 256 total deaths would be required. With an accrual rate of 100 patients per year and a follow-up period of 3 years after the end of accrual, accrual for 4.25 years with entry of 425 patients is necessary.

In general for exponential distributions, the power for detecting a ratio of medians Δ is given by:

$$\text{POWER} = \Phi^{-1}\left\{ \frac{\ln \Delta}{\sqrt{(1/D_1 + 1/D_2)}} - Z_\alpha \right\}$$

where

$$D_1 = (nA/2)\left\{ 1 - \frac{e^{-\lambda_1 F}}{\lambda_1 A}\left(1 - e^{-\lambda_1 A}\right) \right\}$$

A and F are the durations of the accrual and follow-up periods, n is the accrual rate, Z_α is the normal deviate corresponding to an α level type 1 error (1.96 for a two-sided .05 level test), and Φ^{-1} denotes the inverse normal distribution function. λ_1 denotes the failure rate (ln2/median) for the control group and $\lambda_2 = \frac{\lambda_1}{\Delta}$ (4).

Patient Subsets

The calculations above ignore the substantial heterogeneity in prognosis and biology seen in many kinds of human cancer. The calculations assume that treatment comparisons will be made only for all randomized patients as a whole. For many diseases, patient and tumor characteristics have more effect on prognosis than treatment (5). With heterogeneous diseases such as primary breast cancer, there is almost always interest in evaluating relative treatment efficacy within subsets defined by variables such as menopausal status, extent of nodal involvement, and presence of hormonal receptors. Unfortunately, clinical trials rarely include enough patients to perform subset analyses reliably. It is common for reports of trials to describe no overall significant difference but then to compare the treatments within many subsets and find one or more with a statistically significant difference. Although such analyses may provide important information, more frequently such claims are never confirmed (6). Unfortunately, the probability that a significant treatment difference within subsets will occur spuriously by chance alone is large and the power for detecting true treatment differences within subsets is small. For example, in comparing two equivalent treatments within 10 disjoint subsets, the probability of obtaining at least one significant difference by chance alone is 40%. If the subsets are of equal size and the trial was designed for an overall 90% power, then the power of treatment comparisons within subsets is only about 18%.

TABLE 2

Ratio of Median Survival	Number of Total Deaths To Observe
1.2	1268
1.3	612
1.4	372
1.5	256
1.6	190
1.7	150
1.8	122
1.9	102
2.0	88

Total number of deaths in both groups to have power .90 for detecting ratio of median survivals. Type 1 error α = .05 (2 sided).

To perform reliable subset analyses, one must use statistical methods to control the probability of false-positive conclusions and must have sample sizes within subsets to provide adequate power for treatment comparisons (7). For stage II breast cancer, this means at least a doubling of the numbers of patients described in the previous section. Studies without adequate numbers within subsets may suggest hypotheses for testing in subsequent studies, but such hypotheses should be expected to be frequently spurious. Without adequate numbers of patients within subsets, the only reliable inference that can be made from individual studies is usually the overall comparison of treatments. For a disease such as breast cancer, this is not satisfactory for major therapeutic questions (eg, chemotherapy vs no chemotherapy) or questions in which treatment by subset interactions seems likely (eg, tamoxifen and hormonal receptors).

Demonstrating Treatment Equivalence

In most studies of a new treatment versus a standard treatment, it is implicit that the new treatment will be widely adopted only if it produces improved survival or disease-free survival. There are an increasing number of studies, however, in which the new treatment is less debilitating and thus will be accepted as long as it does not have inferior efficacy. With a finite number of patients, one can never demonstrate that two treatments are precisely equivalent with regard to outcome, but one can bound the size of difference which might exist. Such studies need to be large because it is important to demonstrate that the loss in survival rate caused by the new therapy is very small at worst. Studies that reliably identify or exclude small differences must be very large. In reporting the results of such studies, it is important to calculate confidence limits for the survival differences that are consistent with the observed data, because significance tests can be very misleading. Confidence intervals are almost always much more informative than significance tests, but the distinction is particularly important for equivalence studies because of the implications of a misinterpretation of nonsignificant differences. Makuch and Simon (8) provided a simple formula for planning the size of such studies based on confidence limit analysis. Improved methods for calculating confidence limits have been more recently described by Santner and Snell (9) and by Mee (10). Although the use of confidence limits for reporting results is very important, sample sizes can be planned using conventional significance testing methods. Assuming that the two treatments are equivalent and that inference will be based on a 90% two-sided confidence interval, table 3 indicates the approximate number of patients per group needed to have a 90% probability of detecting 5% and 10% differences in survival rate (1). As can be seen from the table, to exclude 10% differences in T-year survival rate, one generally requires 300–500 patients per treatment. Excluding 5% differences generally requires at least 1000 patients per treatment. It is assumed that all patients are followed for at least T years.

Interim Analysis

In addition to having a target sample size large enough to detect the treatment differences of clinical significance, it is also important to perform some interim moni-

TABLE 3

Survival Proportion In Control Group	Difference in Survival Proportions	
	.05	.10
.95	503	165
.90	782	232
.80	1232	338
.70	1538	409
.60	1762	445
.50	1730	445

Number of patients per treatment for showing equivalence in order to have 90% probability for detecting differences of 5% or 10%.

toring of results. In some circumstances, the treatment differences are larger than anticipated and accrual can be reliably terminated early. In general, however, reliable interim decision making is complex and large early differences frequently disappear with further follow-up. Many unconfirmable positive reports are probably attributable to inappropriate early termination of studies and premature reporting of results. This was well illustrated in a computer simulation by Fleming et al (11), who assumed that patients were accrued to a two-armed randomized study at a constant rate of 40 patients per year over 3 years. The endpoint was survival and the treatments were equivalent. If analyses were performed every 3 months, then the probability of obtaining a statistically significant result by chance alone at some point during the 4-year trial was 20%. This illustrates how misleading interim results can be if not analyzed properly. The presentation of interim results may stimulate interest in the trial, but it may also hurt accrual because of the presence or absence of differences. Such presentations may also provide unreliable information to the general medical community.

In some circumstances, interim results indicate that the new treatment is very unlikely to be better than control by a clinically significant amount. In such a case the study can be terminated early. This type of interim analysis can frequently be very important but it must be performed carefully with appropriate statistical methodology. One approach to this analysis is described by Ellenberg (12) in this volume. Another method, due to Lan et al (13), is based on the notion of conditional power. At each interim analysis, one calculates the probability of rejecting the null hypothesis if the alternative hypothesis is true and the trial were continued to its target sample size. This power is calculated given the interim results already observed. If the conditional power becomes too small, the trial is terminated. Another approach to early termination without rejection of the null hypothesis is described by Fleming et al (14).

In addition to using proper statistical methods, interim decision making must be based upon data that have been uniformly updated and carefully reviewed. Even under these conditions, interim results may be transient. It is therefore important that published reports state the target sample size, the history of interim analyses, and the reason for terminating the trial (15).

Epidemiology of Clinical Trials

Table 4 illustrates one implication of performing many trials of small size (16). Suppose that 1000 clinical trials of various types are conducted and that in 100 there is a true treatment difference but in 900 the treatments are equivalent. Suppose that statistical significance is declared 5% of the time when the treatments are equivalent and that the power is 90% for detecting treatment differences that actually do exist. Then, on the average, outcomes will be as shown. There will be 45 trials in which statistical significance is declared when the treatments are really equivalent (false positives) and 90 trials in which true differences come out statistically significant (true positives). Thus, of the 135 statistically significant differences, 45 or 33.3% do not represent real treatment benefits. These results depend upon our assumption that only 10% of the trials have true treatment differences of the magnitude designed to detect and the situation is improved if the 10% figure can be increased. The 10% figure is, however, probably realistic if the differences that the trials were designed to detect are implausibly large. The situation can also be improved by declaring differences significant only if the P value is $< .01$. This requires larger sample sizes if power is to be maintained. This analysis does not necessarily imply that only very large studies should be performed. There are certainly advantages to conducting moderate-size innovative trials with the attempt to confirm those that are positive by doing larger trials (16). This strategy is limited, however, by the lack of power at the first stage for detecting treatment effects of moderate size and reluctance to confirm a positive trial.

CONCLUSION

As a result of the revolution in tumor biology, it seems almost certain that dramatic treatment improvements will be developed for some kinds of cancer without the need for large-scale clinical trials. But there is also the need to make reliable stepwise improvements that pave the way for major advances and to conduct studies that identify improvements which may be less than dramatic but are of major public health value. History has also demonstrated the importance of major randomized trials for preventing the adoption of ineffective therapies that were represented as breakthroughs. Performing such trials reliably tends to require larger numbers of patients than are frequently used. For example, suppose we wish to detect a 1-year increase in median survival for a disease with a median survival of 3 years and to reliably do separate evaluations in two subsets of patients. If patients are to be accrued for 4 years and followed for another 3 years before analysis (a 7-year study), then about 1600 patients are required.

If major trials are to be larger, then the criteria for selecting the most important questions to ask must be sharpened. In some cases there are really not an excessive number of promising hypotheses to test. But this will change, and improved methods of performing pilot evaluations, such as use of selection theory designs (17,18) or response surface designs (19), are needed. Improving the efficiency of trials through study of prognostic factors and use of proper sequential analysis methods is important. We also need to continue to evaluate all aspects of

TABLE 4

True Relationship Of Treatments	Number Of Trials	Probability Of Rejecting Null Hypothesis	Expected Results
Equivalent	900	.05	855 True Negatives 45 False Positives
Not Equivalent	100	.90	90 True Positives 10 False Negatives

$$\frac{\text{False Positives}}{\text{Positives}} = 45/135 = 33.3\%$$

Expected results of 1000 clinical trials. It is assumed that 10% of the trials compare treatments different enough in efficacy that there is 90% power for detection. 33% of the statistically significant differences are false positives.

the way we conduct trials in order to identify improvements and simplifications which facilitate the entry of larger numbers of patients in a manner consistent with high-quality clinical research.

REFERENCES

1. CASAGRANDE JJ, PIKE MC, and SMITH PG. The power function of the "exact" test for comparing two binomial distributions. Appl Stat 27:176-201, 1978.
2. YUSUF S, COLLINS R, and PETO R. Why do we need some large, simple randomized trials? Proceedings of the Workshop on Evaluation of Therapy. Stat Med 3:409-420, 1984.
3. GEORGE SL, and DESU MM. Planning the size and duration of a clinical trial studying the time to some critical event. J Chronic Dis 27:15-24, 1974.
4. RUBINSTEIN LV, GAIL MH, and SANTNER TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. J Chronic Dis 36:469-474, 1981.
5. SIMON R. Importance of prognostic factors in cancer clinical trials. Cancer Treat Rep 68:185-192, 1984.
6. SIMON R. Patient subsets and variation in therapeutic efficacy. Br J Clin Pharmacol 14:473-482, 1982.
7. GAIL M, and SIMON R. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics, in press.
8. MAKUCH R, and SIMON R. Sample size requirements for evaluating a conservative therapy. Cancer Treat Rep 62:1037-1040, 1978.
9. SANTNER TJ, and SNELL MA. Small-sample confidence intervals for P_1-P_2 and P_1/P_2 in 2×2 contingency tables. J Am Stat Assoc 75:386-394, 1980.
10. MEE RW. Confidence bounds for the difference between two probabilities. Biometrics, in press.
11. FLEMING TR, GREEN SJ, and HARRINGTON DP. Considerations for monitoring and evaluating treatment effects in clinical trials. Controlled Clin Trials 5:55-66, 1984.
12. ELLENBERG SS, and EISENBERGER MA. An efficient design for phase III studies of combination chemotherapies. Cancer Treat Rep 69:1147-1152, 1985.
13. LAN GKK, SIMON R, and HALPERIN M. Stochastically curtailed tests in long-term clinical trials. Communications in Statistics — Sequential Analysis 1:207-219, 1982.
14. FLEMING TR, GREEN SJ, and HARRINGTON DP. Designs for group sequential tests. Controlled Clin Trials 5:348-361, 1984.
15. SIMON R, and WITTES RE. Methodologic guidelines for reports of clinical trials. Cancer Treat Rep 69:1-3, 1985.
16. SIMON R. Randomized clinical trials and research strategy. Cancer Treat Rep 66:1083-1087, 1982.
17. SIMON R, WITTES RE, and ELLENBERG SS. Randomized phase II clinical trials. Cancer Treat Rep, in press.
18. SIMON R. An integrated design for the selection and testing of experimental therapies in phase III clinical trials. Presentation to Cancer and Acute Leukemia Group B, Cambridge, MA, Apr 1985.
19. CARTER WH, JR, WAMPLER GL, and STABLEIN DM. Regression Analysis of Survival Data in Cancer Chemotherapy. New York, Marcel Dekker Inc, 1982.

DISCUSSION

Q: (Dr. Susan Ellenberg, NCI) Rich, do you think, in general, that the sample sizes that we use for our trials are not large enough—that we're looking for differences that are unrealistically large?

A: (Dr. Simon) Many trials do look for differences that are unrealistically large based on the intervention studied. In some cases the difference is not unrealistically large but the kinds of patients included are so heterogeneous that at the end you're going to want to look at subsets. You probably ought to be planning large sample sizes on that basis.

The Doctor's Dilemma: Physician Participation in Randomized Clinical Trials

Kathryn M. Taylor[1,*]

Randomized clinical trials are used when the effectiveness of a particular treatment is in doubt or when controversy surrounds the treatment of choice (1). Such trials are experiments, which, in their simplest form, randomly allocate consenting patients to either the traditional treatment (known as standard treatment of choice) or an experimental one. The results are then compared. Each patient in the clinical trial has an equal chance of receiving either treatment. These trials are most common in diseases such as breast cancer, where there is often little scientific evidence to recommend one treatment over another (2). To compare a new treatment with the standard one, two identical patient populations simultaneously receive one or the other of these treatments. Some statisticians argue that randomizing therapy is the best way to ensure valid and reliable comparisons (3). Consequently, randomized clinical trials have become a popular method for testing new medical treatments. However, the overwhelming reluctance of physicians to put their eligible patients on randomized clinical trials is believed by many researchers to be the major impediment to the successful completion of such experimental programs (4,5).

There appear to be two distinct philosophies, either experimenter or therapist, among physicians with regard to participation in clinical trials. Six major characteristics distinguish one philosophy from the other. The first characteristic is the physician's perception of primary allegiance to the individual patient or to the aggregate and future patients. The second difference centers on the physician's interpretation of medical uncertainty. Some physicians anticipate the uncertainty inherent in most clinical decision making and choose to act, rather than refrain from acting, when the treatment of choice is less uncertain. Others, however, interpret the uncertainties in the choice of optimal treatment as a signal to begin experimental procedures to provide a definitive answer. The third difference focuses on physicians' willingness to share information with, or to withhold it from, patients. Some physicians prefer to keep discussions of uncertainty within professional confines. Others, however, argue that patients should be told about a medical controversy and encouraged to participate in decision making. A fourth divergence relates to the physician's views of the reliability and validity of personal clinical experience. Some physicians regard clinical experience as a critical part of any medical decision, while others discard personal experience as anecdotal information of little value, preferring to rely solely on published scientific data. The fifth difference is the extent of loyalty to current therapy. Some physicians adopt a conservative, traditional approach to medical practice, while others, in sharp contrast, question every procedure. The final difference relates to the physician's relationship with colleagues. Some doctors continue an individualized approach to patient care, while others are anxious to pool information resources. It appears that the decision to enter, or not to enter, eligible patients onto clinical trials is strongly related to these six factors.

A recent study clearly illustrates this dilemma (6). In 1980, the National Surgical Adjuvant Project for Breast and Bowel Cancers, the largest North American cooperative group dedicated to research in breast cancer, was seriously concerned that its current clinical trial, B-06, would not achieve successful completion owing to the small numbers of eligible patients being placed on the trial. B-06 was a randomized clinical trial initiated to compare treatment outcomes for patients with breast cancer randomly assigned to one of three treatment groups: segmental mastectomy without radiation; segmental mastectomy with radiation; or total mastectomy. To examine factors influencing the low accrual rate, the principal investigators at all 94 member institutions were asked to complete a questionnaire. A 97% response rate after a single mailing was achieved. The sample frame of principal investigators included 89% from the United States and 11% from Canada. The institutions ranged from large teaching hospitals (59) to smaller, community-based institutions (32). The questionnaire asked whether the surgeons entered all, some, or none of

[1]Department of Behavioral Science, University of Toronto, and Department of Medicine, Women's College Hospital, Toronto, Canada.

[*]*Reprint requests to:* Kathryn M. Taylor, PhD, Department of Behavioral Science, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada M5S 1A8.

their eligible patients into the trial; it also asked surgeons to comment on the scientific design of the trial and addressed the issue of informed consent as a possible obstacle to accrual. Space was provided for open-ended in addition to closed-ended responses. The results were rather surprising.

Over 33% of the principal investigators had not entered any patients into the trial, and 38% stated that they had enrolled only a portion of their eligible patient population. Obtaining informed consent was noted as a difficulty by 25% of the respondents. In an analysis of both open-ended and closed-ended responses there appeared to be six distinct, though interrelated, obstacles to the enrollment of patients in B-06.

1. *Doctor-patient relationship.*—Seventy-three percent of the physicians who did not enroll all patients made some reference to their relationship with the patient. Participation in a clinical trial has been described as potentially compromising of the authority of the physician, and it is not surprising that such capitulation placed pressure on the doctor-patient relationship. Relationships between doctor and patient within a randomized clinical trial are distinct from traditional medical practice, since the ultimate decision-making power is removed from either physician or patient. It appeared that concern about this shift had a major impact on physicians' reluctance to enroll patients in the trial.

2. *Informed consent.*—In addition to other reasons for nonenrollment, 38% of the respondents cited trouble with informed consent. Enrollment of a patient in a randomized trial suggests that the patient has signed a document confirming her understanding of the disease, its risks, and possible side effects of any therapy and has agreed to be randomly assigned to treatment. The respondents appeared concerned that the informed consent form regulates formerly flexible decisions. They expressed an increased discomfort in having to disclose this information to patients. Some physicians (18%) noted that they systematically refrained from attempts to interpret the patient's willingness to hear information since they could no longer act on their personal judgments. The restriction of the informed consent form appeared to be a significant variable in influencing their decision not to enroll patients in B-06. The data also suggest that if the requirement for informed consent were removed, surgeons might still not enter patients in the trial. Furthermore, it appeared that physicians may be differentially selecting patients they feel would not have difficulty with the informed consent document.

3. *Discussing uncertainty.*—Twenty-three percent of the physicians expressed difficulty in telling patients that they did not know which surgical procedure was better. Although they admitted that similar uncertainty exists for surgeons not entering patients in clinical trials, they were unaccustomed to including the patient in frank discussions about the controversy over treatment choice.

4. *Physician as clinician versus physician as scientist.*—Eighteen percent of the physicians argued that they were pragmatists, preferring to rely on their clinical judgment, even when it conflicted with published data. This attitude led them to believe that they had a priori knowledge about which operation was correct, thus obviating the need for scientific investigation.

5. *Practical difficulties in trial procedures.*—In discussions of barriers to accrual, some physicians were concerned about the additional time required to follow a patient enrolled in a trial. Nine percent of the physicians argued that the time required to explain procedures to patients, the rigid rules governing eligibility, the inflexibility of prescribed treatment formulas—all designed to ensure scientific rigor and generalizability of results—were insurmountable barriers to accrual.

6. *Personal responsibility.*—A final obstacle to entering patients in the trial was the fear of personal responsibility should one treatment be more successful than the other. Eight percent of respondents were concerned that should either lumpectomy or mastectomy be superior, personal responsibility would be felt for patients randomly assigned to the arm that was ultimately shown to be inferrior.

It appears from this study that physicians who responded to the questionnaire had adopted a distinct philosophy: a set of integrated attitudes and perspectives which determined their subsequent behavior. These attitudes and perspectives reflected either an experimenter or a therapist philosophy. The experimenters were a small percentage (27%) of the physicians. They claimed primary allegiance to obtaining scientific data and were strong supporters of randomized clinical trials. They felt that primary responsibility was to the aggregate—to the large group of patients to come—and not to the individual patient. They emphasized the rational and technical dimensions of their practice and felt that optimal patient care would be an outgrowth of their research interests. They questioned any unproved treatments based on medical tradition, preferred to share decision-making responsibilities with patients, and interpreted the loss of decision-making power with any randomized clinical trial as a necessity of pure scientific research. In sharp contrast to these physicians, the majority (73%) were therapists, more traditional physicians who argued that they had to take a stance on an issue, despite the controversial nature of their work. They based their decisions on the individuality rather than the general features of each case, and clinical experience was argued to be of primary importance. They interpreted their primary allegiance to the individual patient rather than the aggregate and preferred to act rather than to refrain from acting in the face of uncertainty. These physicians preferred to keep uncertainty in the "backroom," arguing that being totally honest was not always in the patient's best interest. They were basically pragmatists, prefer-

ring to rely on their own conclusions, and were more willing to accept the therapeutic limitations of current treatments.

In conclusion, this study appears to support the belief that the physician's treatment philosophy may be a significant variable in the decision to enter or not to enter an eligible patient onto randomized clinical trials. It also suggests that this hypothesis may prove to be useful if tested on a larger number of physicians in a variety of medical settings.

REFERENCES

1. CHALMERS TC. The impact of controlled trials on the practice of medicine. Mt Sinai J Med (NY) 41:453–459, 1974.
2. FISHER B. Cooperative clinical trials in primary breast cancer: a critical appraisal. Cancer 31:1271–1286, 1973.
3. ZELEN M. A new design for randomized clinical trials. N Engl J Med 300:1242–1245, 1979.
4. LACHER MJ. Physicians and patients as obstacles to a randomized clinical trial. Semin Oncol 8:424–429, 1981.
5. MCCUSKER I, WAX A, and BENNETT JM. Cancer patient accessions into clinical trials: a pilot investigation into some patient and physician determinants of entry. Am J Clin Oncol 5:227–236, 1982.
6. TAYLOR KM, MARGOLESE RG, and SOSKOLNE CL. Physicians' reasons for not entering eligible patients in a randomized clinical trial of surgery for breast cancer. N Engl J Med 310:1363–1367, 1984.

DISCUSSION

Q: Dr. Marvin Zelen has suggested that it's possible and statistically defensible to randomize patients before they go through the informed consent process, then offer the patient that particular treatment group. This may avoid some of the problems that clinicians face in randomization discussions. I'd like to hear some comments about the validity of that approach.

A: (Dr. Taylor) Prerandomization, which was implemented in study B-06 after this questionnaire was completed, increased the accrual rate about 400%. It seemed to me that it made a lot of sense, because what in fact had happened was that the physician was returning to a more traditional relationship with his patient, where he offered options but in fact knew to which arm the patient had been randomized. And it's only conjecture, but I feel that that was an important part of why it worked in this trial. Would somebody else like to comment on the statistical implications of prerandomization?

(Dr. Cyrus Mehta, Harvard) The approach is statistically valid as long as you compare the two treatment groups on the basis of the treatments that they were assigned, not the treatments that they actually got. The difficulty is that there's a dilution effect because a number of patients who were intended to get surgery plus adjuvant therapy might switch over and refuse to take the adjuvant therapy and, conversely, a number of patients who were supposed to get only the surgery would insist on switching over and getting the adjuvant therapy as well. In that case, because of that dilution effect, you're going to need a very large number of patients to compensate and still have valid conclusions at the end. We've had this problem with one of our clinical trials in the Eastern Cooperative Oncology Group (ECOG), where about 25% of the patients switched over from each arm, requiring modified sample sizes which were unrealistic.

Zelen has pointed out that if the comparison really doesn't make good clinical sense, then you'll find tremendous dilution. But if it's a reasonable comparison, there won't be very much of a dilution effect.

Q: (Dr. Bradford Patterson) I have to ask Dr. Taylor if she really feels this particular study is generalizable at all. It worries me enormously that we will draw any conclusions about the implications of physicians' participation in trials from this one, since the B-06 trial challenged surgical thinking, surgical tradition, and ethical issues in a way that almost no other trials which I know about have done or may do in the future.

A: I've given this paper a number of times and had some response, not only from people who were involved in this trial but also from people who were involved in other trials, saying that although this had to do with breast cancer, really you were talking about issues that I felt were important. This is totally anecdotal; I have no way, at the moment, of knowing how generalizable. It is possible that these feelings were elicited only by this trial with two such different arms and will never be seen again. In order to answer that we're now planning to do a prospective study of physician decision-making. When the data come in from that, I think we'll have a little better idea.

Q: Do you have any comments about the informed consent form that was used? Was it intimidating? Was there something wrong with it? Any particular comments on it?

A: I have a sense that the form may be a smokescreen, because it appeared from the analysis that even those who didn't have trouble with the consent form were not entering patients. In our new prospective study one of the variables we're going to be looking at is the informed consent. We're hoping to do a Canadian study along with an American study, looking at differences in the length of the informed consent sheet, the legal requirements, etc. I'm not sure that informed consent is the issue that we think it is. I think the problem is a lot more complex and we've only started to scratch the surface.

Q: I wonder whether there's any difference in the response among physicians who are mainly community-based versus those who are institution-based.

A: That's another really important question. In the new study, that's one of the variables that we're going to look at. We want to find out if there is such a thing as an institutional policy toward clinical trials and, if so, whether it is different in a large urban teaching hospital than it is in a community hospital. Another difference is the physician's relationship with the referring physician. One of the hypotheses is that, in a community hospital, the physician may be more dependent on referrals from people who may be less supportive of clinical trials than a physician in a tertiary hospital whose reputation may be based more on publications and who may have less interaction with the direct referring physician.

Q: (Dr. Jim Anderson, Boston) It's usually accepted that informed consent should consist of at least three things: notifying people that the therapy is research, that it involves randomization if randomization is a part of it, and that the relative benefits of the treatment under study are unknown. If that was part of the process when prerandomization was introduced, what do you think changed to allow for such a dramatic increase in accrual?

A: It seems to me, as a sociologist, that the prerandomization option would be more comfortable for a physician when you compare this to the traditional relationship in which patients come with their symptoms and physicians describe the options and choose one. It's only conjecture, but I feel that the closer we get in organizing clinical trials in ways that are scientifically rigorous and yet are closer to the traditional doctor-patient relationship, the higher the acceptance among the physicians will be of the whole clinical trials process.

Policies for Interim Analysis and Interim Reporting of Results

Judith Rich O'Fallon[1,*]

The key ethical issues involved in developing policies for interim reports and interim analyses are considered. Then the twin topics of the appropriate contents of interim reports and the components of an appropriate interim analysis are discussed. Finally, suggested policies are offered. [Cancer Treat Rep 69:1101–1103, 1985]

As a participant or observer in many discussions about the appropriate content of regular reports for clinical trials conducted by cooperative groups and about the appropriate content and frequency of interim analyses for clinical trials conducted by single institutions and cooperative groups, I have been struck by the fact that such discussions invariably develop into arguments about the proper application of ethical principles: "It's not right to withhold study results from the participating physicians; after all, they've worked so hard to enter patients into this study that they deserve some reward for their effort!" "Yes, but what happens if one treatment arm begins to look bad? Will they feel ethically bound to stop entering their patients into the study to protect them from what they perceive as an inferior treatment?" "Is it right to stop accruing patients before there is adequate evidence that one treatment really is inferior to another?" "Isn't it unethical to waste patients and treatment resources on a nondefinitive study?"

Because the twin topics of interim analyses and interim reports are, in my experience, so intrinsically mixed with ethical problems, I would like to approach these topics by considering the key ethical issues involved. Let us begin with a principle with which, I hope, everyone can agree.

PRINCIPLE: It is unethical *not* to monitor the study for signs of undue risk to participating patients.

In practice, this means that it is necessary to generate the observed distributions of toxic effects, tumor shrinkage, disease progression times, and survival times on a regular basis, eg, every 4–8 months depending on accrual, and to terminate an arm or even the entire study if there is good evidence of substantial damage to patients.

Now let us consider some of the major ethical problems which readily occur to those who are given interim treatment efficacy information:

1. If response/survival distributions *differ substantially*, the following questions arise in the minds of the beholders:
   a. Is it ethical to keep the inferior arm open for patient accrual?
   b. Is it ethical to continue treating the patients already in the inferior arm?
   c. Is it ethical to continue to enter (my) patients into this study as long as the inferior arm remains open?
2. If, on the other hand, the response/survival distributions *are nearly identical*, other questions occur to those who observe these results:
   a. Is it ethical to continue entering patients into this study when one treatment is much more toxic than the other(s)?
   b. Is it ethical to continue entering patients into this study when one treatment is much more costly in time or money?
   c. Is it right to "waste" patients on a negative study, ie, to continue accrual for statistical purposes when more interesting treatments are waiting to be studied?

As soon as serious consideration is given to early termination of a treatment arm or the whole study, however, more ethical dilemmas arise:

1. When a study is stopped or an arm is closed before the planned accrual has been achieved, the study immediately loses the power to detect differences between the treatments that were originally (ie, when the study was designed) considered important to find.
2. The ability to assess long-term treatment damage (eg, late-occurring toxic effects, treatment-induced malignancies) will be decreased because the smaller samples provide less opportunity to observe the occurrence of relatively rare events.

[1]Cancer Center Statistics Unit, Mayo Comprehensive Cancer Center, Rochester, MN.

*Reprint requests to: Judith Rich O'Fallon, Cancer Center Statistics Unit, E12A, Mayo Comprehensive Cancer Center, Rochester, MN 55905.

3. The ability to perform important subgroup analyses will be decreased or even lost entirely.
4. Mistaken early judgments about treatment efficacy could permit thousands of future patients to receive a relatively ineffective, costly, or toxic treatment (if we erroneously report that an inferior treatment is superior) or to fail to receive a relatively effective treatment (if we report that a superior treatment is no better than the current standard therapies).
5. Valuable scientific resources (eg, study patients, research funds, precious supplies of drugs or time in expensive treatment facilities, researchers' time and effort) could be wasted in a noninformative or misleading study.
6. Estimators used to analyze study results (eg, confidence interval estimators) could become biased because of the decision rules used to stop the study early.

On the other hand, if we don't stop early, and final results eventually prove to agree with the preliminary results from interim analyses, then dozens (hundreds) of study participants have been subjected to ineffective, costly, or toxic treatment.

In short, the conduct of clinical trials involves painful tensions between our responsibilities for the welfare of today's patients and our responsibilities for the welfare of future generations of patients just like them.

The ethical dilemmas discussed above almost inevitably suggest themselves to anyone who sees the results of an interim analysis. It is an unfortunate fact that the response tables and survival/progression curves in the traditional routine reports issued by cooperative groups contain enough treatment efficacy information to raise the ethical questions discussed above in the mind of the beholder. Thus, they constitute a series of interim analyses for ongoing studies.

Furthermore, blinding the treatment arms in the reports does not prevent ethical problems from arising. It has been common practice to disguise the identities of the treatment arms in reports by devices such as using different letters to identify the curves and response distributions for the treatment groups. This really accomplishes nothing. If the response percentages or survival (progression) curves are virtually identical for all arms, the results are fully known. This raises questions about the ethics of accrual into toxic or costly arms, wasting patients/resources on a negative study, and continuing accrual for statistical purposes. If the response/survival distributions are very different, participating investigators get nervous about keeping the inferior arm open, continuing to treat patients already entered into the inferior treatment group, and entering their own patients into a study in which they have a substantial chance of getting apparently inferior treatment. Thus, patient accrual may be quietly turned off at the grassroots level.

The danger in widespread distribution of the usual routine reports generated for clinical trials is that such reports constitute very superficial interim analyses, whereas proper interim analyses normally are quite complex, generating interrelated results that must be interpreted cautiously. Major questions that should be addressed in a proper interim analysis include the following.

1. Are treatment differences being exaggerated or hidden by:
   a. imbalances in the baseline characteristics of the groups? (Is one group composed of patients with more favorable prognostic factors?)
   b. differential delays in the submission of evaluation data? (For example, in some multimodality treatment studies, initial treatment and evaluation data for, say, the radiation therapy arms arrive at the statistical office much later than the corresponding data for the surgery or chemotherapy arms because they have to go through special radiation therapy quality control processes first.)
   c. unexpected biases in patient evaluation procedures? (For example, in a randomized study of chemotherapy by implantable pump versus iv chemotherapy, it could happen that surgeons instinctively cancel patients in whom they were unable to implant the pump for various physical reasons. Since the iv chemotherapy group are not similarly ruled ineligible by their physicians following randomization, there are systematic biases introduced into the evaluation data for the two treatment groups that must be addressed before the treatment comparison can be properly interpreted.)
   d. different treatment outcomes in different subsets of patients? (It sometimes happens that treatment A is superior to treatment B in one subset of patients, while treatment B is better than treatment A in the remaining patients. In such cases, the two treatments may appear to be virtually identical in the total patient group.)
2. When does a high P value (say, $P > 0.10$) constitute strong evidence that there are no real treatment differences? To interpret negative results properly, there are a couple of questions that should be addressed:
   a. Has the most powerful test for the problem been used? (For example, in comparing distributions of a variable such as toxicity severity which has values that can be ordered from smallest to largest, it often happens that the chi-squared test is applied when another test such as the Wilcoxon test would have more power to detect meaningful differences.)
   b. Have covariate analyses been done to assess the association of treatment with the responses or

survival variable after adjustment for the effects of other important variables which might not be evenly distributed among the treatment groups?

3. When does a low P value (say, $P < 0.05$) constitute strong evidence that there are real differences among the treatments? To interpret positive results properly, consideration must be given to the following facts:
   a. Frequent repetition of a single test (eg, the log-rank test for the equality of survival distributions) has a very high chance of generating a P value < 0.05 at some time during the trial even when there is no difference among the distributions being compared. In a study done by Fleming et al (1), survival times for two patient groups of size 60 each were generated from the same exponential survival distribution and compared by means of the log-rank test at various times during a simulated clinical trial designed to accrue patients for 3 years and undergo final analysis 1 year after accrual stopped. The investigators repeated their simulated clinical trial 100 times and found that:
      - the log-rank P value was < 0.05 at the final analysis (ie, at Year 4) in exactly five of their 100 clinical trials. That is, applying the log-rank test exactly once—at the final analysis—produced false-positive results in exactly 5% of their clinical trials, as called for in the study design.
      - the log-rank P value was < 0.05 at either the Year 2 interim analysis or the Year 4 final analysis in ten of their 100 clinical trials; ie, applying the log-rank test twice during the trial produced false-positive results in 10% of their trials.
      - the log-rank P value was < 0.05 at least once during the four annual analyses performed at the end of Years 1-4 in 17 of their 100 clinical trials, giving a false-positive rate of 17%.
      - a log-rank P value < 0.05 was observed at least once during the eight semiannual analyses in 21 of their 100 clinical trials, giving a false-positive rate of 21%.
      - a log-rank P value < 0.05 was observed at least once during the 16 quarterly analyses in 26 of the 100 trials, for a false-positive rate of 26%.

      Because this phenomenon has been known to clinical trials investigators for some time, considerable work has already been done to develop appropriate methods for interpreting a series of interim analyses.
   b. Testing several different outcome variables (eg, survival, tumor decrease, time to disease progression, incidence of special toxic effects, various quality-of-life measures) at one point in time has a very high chance of generating at least one P value < 0.05 even when there are no real differences in the distributions of any of the variables among the treatment groups being compared.

Because proper interim analyses are so complex, they should be conducted by a multidisciplinary team of investigators (eg, statisticians and physicians of all relevant specialties) with the skills, time, and patience to evaluate many outcome variables and the relationships among them.

In particular, the monitoring team must be knowledgeable about the expanding methodology for designing and analyzing clinical trials so that they can deal with the subtleties of factor interactions and multiple testing. They must also have the patience to evaluate the consistency of results within patient subsets, between different variables, and relative to results of relevant studies done elsewhere.

SUGGESTED POLICIES

As a consequence of the thought processes discussed above, my clinical and statistical colleagues have concluded that it is wise to adopt the following policies:

1. Treatment efficacy information should not be given in routine general reports distributed to all study investigators until after patient accrual ends. It should, however, be reported regularly to a monitoring committee responsible for the ethical conduct of the trial.
2. Each clinical trial should have a multidisciplinary monitoring committee responsible for:
   a. studying routine reports containing treatment efficacy information
   b. performing appropriate interim analyses
   c. deciding when early termination (arm or study) should be considered
   d. presenting the case for early termination to the whole group of investigators when necessary.

Many experienced investigators associated with cancer clinical trials have expressed skepticism about the feasibility of implementing such policies, citing primarily their doubts about the willingness of clinicians to participate in studies in which the treatment results will not be known for years.

However, our cooperative group implemented these policies more than 5 years ago and has been operating quite successfully—and contentedly—within these constraints ever since, proving that it can be done.

REFERENCE

1. FLEMING TR, GREEN SJ, and HARRINGTON DP. Considerations for monitoring and evaluating treatment effects in clinical trials. Controlled Clin Trials 5:55-66, 1984.

DISCUSSION

Q: Judy, you mentioned keeping the covers on until accrual stops, but patients may actually be receiving treatment for another 2 years after accrual stops, and premature disclosure of the results may lead to people changing some of their treatment policies, so one can make a case for keeping the covers on even longer.

A: (Dr. O'Fallon) That's true.

Q: (Dr. Cyrus Mehta, Harvard) I have two comments. One concerns this idea that all the efficacy results should be kept blinded except by this monitoring committee. I wanted to know who would be on this monitoring committee—would it be a committee of statisticians and physicians? Because if it had physicians on it and the physicians were putting patients on the study, then they would be facing the same types of issues that you mentioned. They would have the choice of holding back their own patients from the study.

A: True, Cyrus, but that problem does not arise, at least not very much, because they are also looking into the data carefully—they are involved in the interim analyses—they know that just a simple glance at these tables or curves is not really sufficient, so they withhold judgment until such time as they begin to feel very strongly that there is something going on.

(Dr. Mehta) I had one other comment: recently, the technologies for doing these types of group sequential tests and early stopping to guarantee a given overall significance level have, in fact, been well worked out, so that's not what's going to hold us back from doing early stopping. What's going to hold us back are all these other issues that you mentioned, like imbalances in the baseline characteristics, unexpected biases in patient evaluations, covariate adjustments, things like that—not the pure technology of doing the group sequential tests. How serious are those other issues, compared to just the mathematics?

A: In order to know whether a separation of the curves means anything, we need to go into all these other issues about the comparability of the groups and procedural problems that we didn't realize we were going to have. As long as the people who are privy to these results are also doing the thinking and the searching for the complications, I think group sequential testing is appropriate, because decisions will not be made only on the basis of a spread between the curves. The monitoring committee will also look at whether these are being caused by structural problems.

Q: (Dr. John McGuire, Triton Biosciences) I appreciate your explanation of the tension that the physician is under and that this is relieved by blinding the investigator to the interim results. But it seems to me that the only time an ethical concern arises is when the state of knowledge about the relative benefits of the treatments differs from what is presented to the patient in the consent form. Since the purpose of the consent form is to tell the patient what the state of knowledge is with respect to the treatments to which the patient is being randomized, how can we ethically give these new patients a consent form which was drawn up prior to these interim results which are now causing some interim tension? Shouldn't the consent form be updated in order to allow the patient to face these issues?

A: The problem is, what do we know? Until we have all the data, until we have all the people in and the data accrued and cleaned up, we don't really know how to interpret the results. And I'm not sure that we have to tell people, well, it might be this and it might be that and it might be something else, in midstream.

Q: (Dr. Bob Birch, University of Alabama) When you have a study which is being monitored by committee, what information is reported in your regular statistical reports, and if it's nothing, how did the NCI accept it?

A: The group receives the administrative statistics, the descriptions of the patient groups, and all the toxicity information. NCI receives a copy of all of the information that goes to the monitoring committees, so in essence, our monitoring committee contains a member from NCI as well.

Q: (Dr. Arnie Weiss, Lederle) Are you using group sequential methods? And if you are, can you tell me which ones you're using and why—Peto versus Pocock versus O'Brien-Fleming.

A: In one-arm studies, the phase II type, we use the Fleming designs and also a variant on them that I call the pessimist's. That's the one by Dr. Green; it isn't published, but in her design, if you get any hope that there's something going on, you keep going. When we get to comparative trials, we use the O'Brien-Fleming design. The choice depends on your philosophy about stopping. That particular stopping rule requires extreme results to stop early. We like that because our patient populations are so heterogeneous that we want to be very conservative in making early decisions to stop. We would prefer to have absolutely overwhelming evidence of treatment differences before we will stop very early.

Q: (Dr. Robert Wittes, NCI) If that's the case, then my question may be irrelevant. From your answer to Dr. Mehta's question before, I wasn't quite clear whether your data-monitoring committees include the various principal investigators.

A: Oh yes, ours do. Our monitoring committee consists of the principal investigator and the statistician, and so on. We're not bringing in outside people as they do in the cardiology trials. We can't afford it.

Q: (Dr. Wittes) You don't feel on the basis of your participation in this over the last 4–5 years that the use of principal investigators in the data-monitoring committee itself is a source of investigative bias?

A: No, we haven't seen anything of that yet, because they are being brought into the complexities of making these decisions. I think they realize that when they are beginning to be really troubled that we are too, and that we should be bringing it to the whole group for consideration of closing.

Q: It seems to me that you have the same problems when you report toxic effects that you have when you report early efficacy results. I agree that you can't blind the toxicity data from the whole group because people need to know what sorts of trouble patients are getting into. I wonder if you have a sense of how much of the problem is taken care of by blinding the efficacy results and reporting the toxicity results.

A: Well, the toxicity results are reported because of patient care considerations. We need to know what types of toxic effects have been experienced. So far as I can tell, we have not had any accrual drying up because of this type of reporting.

Q: (Dr. Robert Gordon, NIH) Just a brief comment. About 5 years ago, the question of whether or not it's ethically imperative to reveal emerging trends to new patients was addressed by the Ethics Advisory Board of the Secretary of what was then the U.S. Department of Health, Education, and Welfare. The Ethics Advisory Board concluded that it would be perfectly ethical to keep interim results under wraps through the duration of a trial if two criteria were satisfied: (*a*) if this was a part of the experimental design that was made explicit at the beginning, so that when the protocol was reviewed by an institutional review board everybody knew that the results were going to be kept under wraps; and (*b*) if the informed consent documents drawn up at the beginning of the trial said to each patient, not only the first, but subsequent patients, that the design of this study requires that interim results be kept confidential. If these two conditions were satisfied there was nothing ethically wrong with keeping the results under wraps all the way through. However, the Freedom of Information Act (FIA) requires that interim results of any clinical trial being done by Federal investigators or by Federal contractors must be made public in response to a Freedom of Information request; this limitation in the FIA still remains unchanged.

Eligibility Exclusions, Losses to Follow-up, Removal of Randomized Patients, and Uncounted Events in Cancer Clinical Trials

Mitchell H. Gail[1,2,*]

This paper reviews the potential impact of eligibility exclusions, losses to follow-up, removal of randomized patients from the analyzed data, and failure to count all observed events on the interpretation of cancer clinical trials. The pragmatic and explanatory approaches to these problems are presented, and it is argued that most clinical trials yield data that are suitable for answering pragmatic questions. Thus, the pragmatic approach to these problems is stressed. [Cancer Treat Rep 69:1107–1112, 1985]

Several recent articles have addressed the handling of eligibility exclusions, losses to follow-up, removal of randomized patients, and uncounted events arising in randomized clinical trials (1–5). This paper illustrates the importance of these ideas for therapeutic cancer trials. Heavy reliance on the literature is apparent, even in the title (1).

Eligibility exclusions, losses to follow-up, removal of randomized patients, and uncounted events are indicated schematically in table 1. Patients who do not satisfy the eligibility criteria for entry in the trial are excluded. Thus, exclusions take place prior to randomization. The other contingencies in table 1 take place after randomization and therefore have the potential to invalidate the comparison between the two randomized treatment groups. The term "loss to follow-up" refers to patients for whom the observation of response following randomization is incomplete. Loss to follow-up takes place after randomization but before analysis. Controllable losses to follow-up, such as those that occur when a patient refuses to continue hospital visits, can be distinguished from uncontrollable losses to follow-up, such as those that occur in a cancer trial in which time to recurrence is the endpoint when a patient dies of heart disease prior to recurrence. The term "removal" will be used to describe the deletion of certain randomized patients from the analysis, for example, on grounds that the treatment was improperly administered. Others (3) have used the term "withdrawal" instead of removal. Uncounted events may arise when the data analyst excludes certain events from analysis, such as recurrences that occur within 2 months of randomization. It is evident from these definitions that the contingencies in table 1 are common and that one must have a policy for dealing with them.

In considering these issues it is helpful to recall the distinction made by Schwartz et al (5) and, earlier, by Schwartz and Lellouch (6) between a pragmatic trial philosophy and an explanatory trial philosophy. The pragmatist accepts the fact that human experimentation is difficult and subject to many constraints that do not apply in the laboratory. He realizes that patients will not always take their scheduled doses of medication and that some will leave the trial altogether. In the pragmatic view, the purpose of the trial is to compare effectiveness of the *intention* to administer Treatment Plan A with the effectiveness of the intention to administer Treatment Plan B, because one can only randomize patients to an intended treatment program, not to a received treatment. Sometimes, one phrases this idea as comparing two treatment policies. In the pragmatic point of view, all of those patients assigned to Treatment Policy A should be compared with all of those assigned to Treatment Policy B, even though some do not receive the full protocol treatments. The pragmatist wants a valid comparison of two treatment policies under realistic clinical conditions.

The explanatory point of view is basically that of the laboratory scientist. Typical explanatory questions are: "How would Treatment A do if all patients adhered perfectly to the protocol dose schedule?" or "What are the biological mechanisms by which Treatment A works?" Even though these are interesting and worthwhile questions, it is by no means clear that they are answerable from data obtained in an imperfect clinical trial setting. Some of the discussion in this paper indicates how at-

[1]Biostatistics Branch, National Cancer Institute, Bethesda, MD.

[2]I thank Dr. David Byar for helpful editorial comments and Ms. Cynthia Fournier for typing the manuscript.

[*] *Reprint requests to:* Dr. Mitchell H. Gail, Biostatistics Branch, National Cancer Institute, Landow Bldg, Rm 3C37, National Institutes of Health, Bethesda, MD 20205.

TABLE 1.—Timing of exclusions, losses to follow-up, removals, and uncounted events

	Description
Eligibility exclusions	Take place before randomization; eligibility criteria affect the composition of the study population
Losses to follow-up	Occur after randomization
Removals of randomized patients	Defined by the investigator at the time of analysis
Uncounted events	Defined by the investigator at the time of analysis

tempts to derive explanatory answers from imperfect cancer clinical trial data can produce misleading interpretations.

Eligibility Exclusions

The eligibility criteria for entry in a trial may affect the *generalizability* of the resulting treatment comparison. However, because the eligibility determinations and exclusions occur prior to randomization, the process of exclusion is not associated with treatment assignment and is therefore not a confounder (7). Thus, regardless of the exclusions, an unbiased treatment comparison between the two randomized groups may be obtained. For this reason, we need only consider the impact of eligibility criteria on the generalizability of the observed result.

One would like assurance that the treatment effect observed in a clinical trial would also be found in general clinical practice. The ideal way to obtain such assurance is to select patients at random from the general population for entry in a given trial; this is impractical. Instead, one hopes that the patients entering a given trial are reasonably representative of the general population of diseased patients, and one relies on the scientific paradigm that if a valid cause and effect relationship has been found in one randomized experiment, the result is likely to hold, at least qualitatively, in other settings. The argument is strengthened if several independent trials in different locations confirm the result.

The eligibility criteria used in cancer clinical trials often reflect a desire to obtain explanatory, clear-cut results. To begin with, trials of a new cancer chemotherapy are often restricted to a single histological cell type and stage of disease. These restrictions tend to produce a fairly homogeneous study population so that the effects of treatment are more easily discernible against the background variability in response. If no such restrictions were imposed, the effect of treatment could be lost in the tremendous variability of response arising from mixed stages of disease and cell types, which are often more powerful determinants of outcome than is the treatment (8). Alternative strategies for dealing with such heterogeneity include stratification on risk factors, covariate adjustment on risk factors, and increasing the sample size to improve the "signal to noise" ratio. These alternatives are employed even in trials restricted to a given cell type and stage to control heterogeneity identified with other important patient characteristics, such as prior weight loss or performance status. If eligible patients are restricted to a single cell type and stage of disease, one obtains no direct empirical evidence that a result found in the present trial will generalize to other cell types or stages of disease, and additional trials are usually required to extend the results.

Other commonly used eligibility criteria may affect generalizability. To improve treatment compliance and follow-up, patients with concurrent diseases and elderly patients are often excluded from cancer chemotherapy trials. Thus, the patients in the trial may follow their prescribed dosage schedules better than patients in general practice, and they may be less subject to complications arising from concurrent disease. Likewise, most cancer trials exclude patients with a prior history of cancer, and pathological specimens are required before randomization. In general clinical practice, it is likely that a higher proportion of putative lung cancer primaries, for example, are in fact metastases from an earlier cancer.

Thus, perfectly valid treatment comparisons for a clinical trial study population defined by explanatory eligibility criteria may not be quantitatively applicable to general practice.

Losses to Follow-up

Table 2 presents a summary of hypothetical survival data after 1 year in a comparison of adjuvant chemotherapy against placebo. Twenty percent of patients are lost to follow-up on each treatment. These losses are called controllable because, in principle, it is possible to retrieve the survival information on these patients by further investigation including, possibly, use of the National Death

TABLE 2.—Controllable losses to follow-up

	Adjuvant chemotherapy			Placebo		
	Followed	Lost	Total	Followed	Lost	Total
Deaths within 1 yr	35 (44%)	15 (75%)	50 (50%)	40 (50%)	10 (50%)	50 (50%)
Survivors	45	5	50	40	10	50
Total	80	20	100	80	20	100

Index. Such losses can also be prevented by stressing the importance of complete follow-up to patients at the time informed consent is obtained. In table 2, the survival data on "lost" patients are also shown, as if a second effort had successfully retrieved these data.

If the 80 patients who were "followed" on each group are compared, it appears that adjuvant chemotherapy is slightly beneficial (44% vs 50% deaths). The question is: Are the two followed groups comparable, apart from treatment? Although the randomization was designed to assure comparability of the entire population on chemotherapy with the entire population receiving placebo, it is possible that sicker patients were lost from the chemotherapy group compared with those lost from the placebo group. This phenomenon, known as "differential loss to follow-up," may induce a bias. In this hypothetical example we have obtained the missing data on the survival of lost patients, and it is clear that the lost patients receiving chemotherapy did worse than the lost patients receiving placebo (75% vs 50% mortality). When we reconstruct the experience of all randomized patients, we find no difference between the two treatment groups, and this is the appropriate comparison. The fact that 80% of the patients were followed on both treatment groups does not guarantee that the two followed groups are comparable; they are not. A more meaningful investigation of comparability of the two followed groups is to compare the distributions of known risk factors in these two groups. Even so, one can never really be sure that the followed groups are comparable, unless loss to follow-up is minimal. Therefore, one should make every effort to prevent controllable losses and to retrieve lost information.

Some losses are uncontrollable. The hypothetical data in table 3 give the numbers of recurrences within 3 years among 80 followed patients receiving chemotherapy and among 90 followed placebo controls. In this experiment, 30 patients died without recurrences within 3 years, and the recurrence data for the full 3 years cannot be determined in these patients; nor can such competing noncancer risks be prevented.

The data on followed patients suggest an advantage for chemotherapy (25% vs 33% recurrences). But as before, the question to be asked is: Are the two followed groups comparable? It is possible that those patients who died of competing risks on chemotherapy were at high risk of recurrence, whereas those who died of competing risks on placebo were at low risk of recurrence. One can put bounds on the proportions with recurrence in the two treatment groups by examining the most extreme possible outcomes for those who died prior to recurrence. A "worst-case" analysis for chemotherapy is to assume that all those patients who died without recurrence on chemotherapy would have developed recurrence, whereas none on placebo would have developed recurrence. In this case, the proportion of recurrences on chemotherapy would be (20 + 20)/100 = 40%, compared to (30 + 0)/100 = 30% on placebo. The corresponding "best case" analysis yields (20 + 0)/100 = 20% recurrences on chemotherapy compared to (30 + 10)/100 = 40% recurrences on placebo. These bounds on performance are similar to those used by Peterson (9) for competing risks when time to response is measured. It is clear that these bounds are too broad to be helpful in this example, because in the "worst case" analysis, chemotherapy seems harmful, whereas in the "best case" analysis, chemotherapy seems beneficial. Whenever the proportion of uncontrollable losses is large, these bounds will be wide. Only if we assume that those followed are representative of the entire original treatment groups can we rely on the simple comparison of the two followed groups, and this assumption is not directly testable.

For these reasons, when there are substantial uncontrollable losses to follow-up, it may be wise to place less reliance on an endpoint like recurrence, and to revert, instead, to an endpoint like survival that is not subject to competing risks. Regardless of the cause of death, it can be determined unequivocally whether or not each patient survived the period of observation. For survival, therefore, the losses are controllable and can be minimized.

Removals

At the time of analysis, there is a strong explanatory temptation to remove some patients from the data set. The two principal justifications for such removals are errors in eligibility determination and failures to comply with the intended treatments. The pragmatic procedure is to analyze all randomized patients, and certainly such an analysis should be included in any publication. Removals may bias the treatment comparison between the two randomized groups, because removals take place after randomization, and, in many cases, there is reason to sus-

TABLE 3.—Uncontrollable losses from competing risks

	Adjuvant chemotherapy			Placebo		
	Followed	Died without recurrence	Total	Followed	Died without recurrence	Total
Recurrences in 3 yrs	20 (25%)	?	?	30 (33%)	?	?
No recurrence	60	?	?	60	?	?
Total	80	20	100	90	10	100

pect that the criteria for removal are related both to treatment assignment and to response.

Patients who do not satisfy the eligibility criteria are often removed from analysis. We shall distinguish three categories of errors in determining eligibility: (*a*) errors on objective data measured before randomization, (*b*) errors on subjective data measured before randomization, and (c) in flawed protocols, eligibility criteria that depend on events following randomization. The first category is typified by an error in age or cell type. As an egregious example, suppose a patient with bronchitis was mistakenly randomized to a lung cancer protocol. We can only hope he also failed to comply with treatment. Should such a patient be removed from analysis? Removing such a patient will not bias the treatment comparison provided the detection of this eligibility violation was unrelated to treatment assignment. To avoid such an association, it is important to use uniform standards on both treatment arms to review for such protocol violations. It is common practice to include an analysis of eligible patients, based on removal of patients with Category *a* eligibility violations of objective pathology, to complement the pragmatic analysis of all randomized patients.

Eligibility violations in Categories *b* and *c* are problematic, and the potential for biased removals is great enough that such analyses should be avoided and must be viewed skeptically. The potential for abuse is especially great if subjective data (Category *b*), such as performance status or histologic grade, are re-evaluated at the time of analysis. Then an overzealous investigator may consciously or unconsciously apply different subjective standards to the two treatment groups, with the result that sicker patients are preferentially removed from one or the other group. The possibility of such bias can be reduced by withholding information on treatment assignments from the reviewers. Category c violations can only arise in a flawed protocol in which eligibility criteria depend, in part, on events following randomization. For example, if any patient who develops a recurrence within 6 months of randomization is removed as "ineligible," one is clearly applying a criterion that is related potentially both to treatment assignment and to ultimate patient response. This is because the absence of recurrence in the first 6 months may be related to an effective treatment and is prognostic of future well-being. Withdrawing such patients may therefore lead to a seriously biased estimate of overall treatment effectiveness. As an extreme hypothetical example, a bad treatment that speeded up all recurrences to fall within 6 months after randomization could appear to be an excellent therapy if patients with early recurrence were removed from the analysis.

Another example of a faulty Category *c* eligibility criterion is the requirement that treatment begin within a fixed time interval following randomization. For example, suppose resected lung cancer patients are randomized to adjuvant chemotherapy or no further treatment, with the stipulation that chemotherapy begin between 30 and 40 days following randomization. Then, all of those patients randomized to no further treatment will be compared with the subset of chemotherapy patients who are well enough to begin treatment in the prescribed time interval. This subset of patients may very well be healthier than the entire set of patients originally randomized to chemotherapy, and a biased estimate of treatment effect may result. This example illustrates the desirability of randomizing patients just before the treatments are to be given, and it leads us to the following discussion of treatment compliance.

The second common justification for removals is for failure to comply with the intended treatment, as when patients refuse chemotherapy following randomization or when leukopenia forces a nonscheduled reduction in dose. The pragmatist accepts such eventualities as part of the intended treatment policy and does not remove such patients. The explanatory investigator seeks to compare only those patients who get full treatment on both treatment groups, as indicated in the hypothetical example in table 4. In that example, there is no advantage for chemotherapy overall, since 50% survive 1 year in both treatment groups. However, among the 60 patients assigned to chemotherapy who received the full dose, 67% survived 1 year, as compared with 50% who received the full treatment on the control arm, namely, no further treatment. As in the Losses to Follow-up section, the question to be asked is whether those patients who received the full treatment on chemotherapy are comparable, apart from treatment, with those who received the full treatment on the control arm.

As previously discussed, randomization only induces comparability of the two entire randomized groups, not necessarily of subsets of these groups. It is plausible that those patients who could tolerate full doses of chemother-

TABLE 4.—Misleading explanatory interpretation

	Chemotherapy			No further treatment		
	Full treatment	Non-complier	Total	Full treatment	Non-complier	Total
Survive 1 yr	40 (67%)	10 (25%)	50 (50%)	50 (50%)	0	50 (50%)
Die	20	30	50	50	0	50
Total	60	40	100	100	0	100

apy are healthier on the average than the entire group initially assigned to chemotherapy. If so, the apparent advantage for chemotherapy within the full treatment subset may be due to selection, rather than to a treatment effect. Thus, it may be reasonable to ask the explanatory question, "What is the effect of chemotherapy if all patients comply fully?" However, we cannot derive a convincing answer from a clinical trial in which a substantial fraction of patients deviate from full compliance. The comparison of those who received full treatment may be completely misleading. For this reason, analyses that remove patients who do not comply should usually not be published, or, if published, should be accompanied by warnings for the reader. Redmond et al (10) discuss the closely related difficulties of trying to determine a dose-response relationship by retrospectively classifying patients according to dose received.

We illustrate these ideas with three additional hypothetical cases. Suppose it is discovered that a patient randomized to a lung cancer protocol had had prior gastrointestinal cancer, an exclusion criterion for the lung cancer protocol. Suppose that this information was written clearly in the medical chart, prior to eligibility determination, and that the error resulted from simple oversight. This patient might be removed for a Category *a* eligibility violation if one could be assured that the detection of the error was not related to which treatment had been assigned. As a second example, suppose that 6 months after randomization to a non-small cell lung cancer protocol the pathology reference center renders an opinion that the patient had a small cell tumor. We assume that the specimen had been sent to the pathology center prior to randomization and that the reference center was not given data on treatment. In these circumstances, it is unlikely that the new information on eligibility exclusion is related to treatment, so removal is unlikely to bias the treatment comparison. Nonetheless, the pragmatist may object to removal on grounds of generalizability. His argument is that the purpose of the trial is to determine the effect of treatment under conditions of clinical practice. But in practice, the physician must make the diagnosis and treatment decision on the basis of local pathology resources, which may not include specialized tests available only at the pathology reference center. Thus, an argument is made that the estimate of treatment effect is more applicable to general practice if the patient is retained in the analysis. As a final example, suppose a patient develops a cardiac arrhythmia that precludes further chemotherapy after only one cycle of treatment. This adverse clinical event may be related both to the treatment assignment, because some agents such as doxorubicin induce arrhythmias, and to prospects for prolonged survival. Thus, removing such a patient may induce bias in the estimated treatment effect, and the pragmatist will retain the patient. Those investigators who remove such a patient in an effort to study the explanatory role of full treatment on chemotherapy may be deluded by results that appear to show a benefit for full treatment with chemotherapy when, in fact, all that has been demonstrated is that those who could tolerate full treatment were hardy to begin with.

To summarize, the publication should include an analysis of all randomized patients. An analysis that removes certain patients for gross Category *a* eligibility violations should also be presented. Other supplementary analyses, based on removal of patients for Category *b* or *c* eligibility violations or for failure to receive full treatment, are not fully protected by the randomization, because the criteria for removal may select out healthier or sicker patients differentially from the two treatment groups. Therefore, if such supplementary analyses are presented, the reader should be warned of possible biases.

Uncounted Events

If a chemotherapy is designed to reduce death from cancer, a natural explanatory analysis is to focus on "death with cancer," which is defined as death in a patient who has evidence of cancer at the time of death. For purposes of this analysis, patients who died of unrelated causes without evidence of cancer, like heart disease, are treated as censored observations at the time of noncancer death. Thus, the events are not counted. The pragmatist, and especially the patient, are more interested in survival from all causes. A potent treatment that kills the tumor but also kills the patient is useless. The analysis that counts all deaths, from every cause, is therefore needed to describe the clinical trial fully. Although the analysis of only "deaths with cancer" may provide information as to whether chemotherapy is having a biological effect, two limitations should be remembered. (*a*) As just mentioned, the explanatory question of biological effect is less important to the patient than the pragmatic question of overall survival. (*b*) The analysis of "death with cancer," like the analysis of recurrence in table 2, is subject to biases from uncontrollable loss to follow-up, because we don't know whether those people who died from "noncancer causes" were at unusually low or high risk of death with cancer when they died. Thus, although it is reasonable to carry out analyses of death with cancer, the analysis of all deaths should also be presented.

Chemoprevention techniques to reduce cancer incidence may not take effect until several years after the treatment begins. This idea may cause the analyst to count only those events that occur after 1 year, as indicated in the hypothetical example of table 5. If one compares the proportions with cancer incident in Years 2–5, it appears that chemoprevention is slightly beneficial (8% vs 9%). This finding may have resulted from a delayed biological effect of chemoprevention, and this analysis may be important as a demonstration of biological effectiveness. However, from the pragmatic point of

TABLE 5.—Uncounted events*

	Chemoprevention	Placebo
Cancer develops in 1 yr	20	10
Cancer develops in 2–5 yrs	80	90
No cancer at 5 yrs	900	900
Total	1000	1000

* Values = No. of patients.

view, it is also important to present the analysis of all cancer cases, which shows that at the end of 5 years 10% of the patients developed cancer on each treatment. Similar considerations apply to therapeutic trials in which it is felt that an early recurrence cannot be prevented by treatment.

To summarize, specific analyses that are intended to demonstrate a biological effect may focus on a subset of the observed events. However, to present the trial results fully, especially with regard to pragmatic impact, an analysis of all events should also be included. Moreover, analyses that focus on a subset of observed events may be subject to biases from uncontrollable losses, namely, the patients whose events are uncounted.

DISCUSSION

In this paper I have pointed out that criteria for excluding ineligible patients before randomization may effect the generalizability of the results but not the validity of the treatment comparison between the randomized groups. Cancer clinical trials employ a number of explanatory eligibility criteria, designed to make the treatment comparison easier to interpret, perhaps at the cost of making it harder to determine whether the results are widely applicable.

Losses to follow-up, removals, and uncounted events all occur following randomization and may lead to biased estimates of treatment effect. The pragmatic approach to these problems is to adopt endpoints like survival that are less subject to uncontrollable losses to follow-up, to analyze all randomized patients as well as a subset containing all but a very narrow class of patients with Category *a* eligibility violations, and to perform an analysis counting all events. These procedures answer a number of pragmatic questions within the framework of the original randomized comparison. One must view attempts to answer explanatory questions cautiously, whenever these methods (*a*) rely on endpoints that are subject to heavy uncontrollable losses, (*b*) remove many patients from the original randomized treatment groups, or (*c*) fail to account for all events.

REFERENCES

1. ARMITAGE P. Exclusions, losses to follow-up, and withdrawals in clinical trials. *In* Clinical Trials, Issues and Approaches (Shapiro SH, and Louis TA, eds). New York, Marcel Dekker, 1983, pp 99–113.
2. DEMETS DL, FRIEDMAN LM, and FURBERG CD. Counting events in clinical trials. N Engl J Med 302:924, 1980.
3. MAY GS, CHIR MS, DEMETS DL, ET AL. The randomized clinical trial: bias in analysis. Circulation 64:669–673, 1981.
4. SACKETT DL, and GENT M. Controversy in counting and attributing events in clinical trials. N Engl J Med 301:1410–1412, 1979.
5. SCHWARTZ D, FLAMANT R, and LELLOUCH J. Clinical Trials. London, Academic Press, 1980.
6. SCHWARTZ D, and LELLOUCH J. Explanatory and pragmatic attitudes in therapeutic trials. J Chronic Dis 20:637–648, 1967.
7. SCHLESSELMAN JJ. Case-Control Studies: Design, Conduct, Analysis. New York, Oxford University Press, 1982, p 58.
8. BYAR DP. The necessity and justification of randomized clinical trials. *In* Controversies in Cancer (Tagnon HJ, and Staquet MJ, eds). New York, Masson, 1979, pp 75–82.
9. PETERSON AV. Bounds for a joint distribution function with fixed sub-distribution functions: application to competing risks. Proc Natl Acad Sci USA 73:11–13, 1976.
10. REDMOND C, FISHER B, and WIEAND HS. The methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease-free survival. Cancer Treat Rep 67:519–526, 1983.

DISCUSSION

Q: (Dr. Joe Pater, Kingston) One problem that cooperative groups face is centers which don't comply or systematically don't administer the treatment as effectively or as aggressively as other centers. What would your approach be to that problem? Or are we to take the same approach? Do both analyses?

A: (Dr. Gail) I think you're obligated to include those results in your presentation of the data and certainly I don't consider that a reason for exclusion from the trial. Presumably the reason you studied many centers was to get some idea of the variability of the treatment effect in a variety of settings. This is a strength as well as a weakness of your study.

Q: You suggested that we should look at whether patients met the eligibility criteria and then decide whether or not to withdraw them. One very common eligibility criterion in our cancer trials is that patients should have adequate BUN, creatinine clearance, bilirubin, and so on. These criteria are primarily set up to protect the patients from getting subsequent serious toxic effects. The trouble is that there are large numbers of patients for whom these values are not well documented or who really failed to meet those criteria but do go on study and don't get severe toxic effects. I was wondering where you would stand on these patients, would you exclude them from the analysis or would you not?

A: Well, unless I thought that the reason you detected failure to have this information was related to subsequent outcomes, I would tend to ignore this sort of problem. I would definitely include those patients in the analysis.

Q: (Dr. Satya Dubey, FDA) Suppose one has performed two kinds of analyses, one with all the patients who were randomized and another with only those patients who completed the studies. At the end, even though there were various reasons for loss to follow-up, the statistical conclusions are very similar. How do you interpret and report these results?

A: The all-randomized analysis is essentially an unbiased analysis. The only question is whether it is an important comparison to be making, because if there is a very substantial number of people who really didn't get the assigned treatment, you may have a perfectly valid comparison, but it may not be telling you very much. The treatment effects may be vastly diluted. If you get the same results when you do that as when you do the analyses closer to the explanatory viewpoint, it is very reassuring. You know on the one hand that the randomized analysis is unbiased, and also it seems to be telling you the same thing as what you might call the relevant analysis.

Measuring Quality of Life: Risks and Benefits[1]

H. Schipper* and M. Levitt[2]

Quality of life research is an emerging science of particular relevance to clinical cancer research. The development and utilization of valid and reliable quality of life measures as outcome parameters may profoundly alter the clinical trials process. Such measures, contrary to prevailing opinion, may be as accurate and precise as conventional measures, but considerably more relevant to the trials process. However, lack of rigor in the evaluation of such indices and uncritical interpretation of results may seriously compromise the credibility of the concept. This overview of the current status of quality of life research is designed to provide a perspective from which to evaluate ongoing development. [Cancer Treat Rep 69:1115–1123, 1985]

Endpoints, Outcomes, Quality of Life

In scientific research, what you see is what you get. The corollary is that if you don't look for it you won't find it.

Recently, some disturbing trends have been observed in cancer research. There is growing skepticism about the benefits of treatment for a number of malignant diseases; patients are more than occasionally refusing what would seem to be "proven" therapies. In the face of disappointing results, ever more toxic therapies are being tested. One senses the ultimate prospect of powerfully effective therapies which patients refuse.

An insight into this conundrum may be offered by something of a return to basic principles of experimental design. First and foremost, the selection of endpoints defines absolutely the range of possible experimental outcomes. In oncology, the emphasis has been on response rates, disease-free survival, and absolute survival (1). These endpoints were selected because they were considered measurable and consequently "scientifically" evaluable. Even though there has been a gradual recognition that these parameters are not as firm as was initially believed, we have continued to depend on them to be reliable signposts to success in the step-wise trial model of clinical cancer research. The net effect is that we design treatment to shrink tumors rather than to preserve and augment human function. Quality of life is a measurable endpoint, more relevant to patients than our traditional outcome measures. The development and application of adequate tools to measure quality of life may substantially enhance the clinical trials process.

Our reluctance to use quality of life measures to assess clinical outcome has historical roots. From Hippocrates to Osler, medicine has been held as an art. Physicians were the mystical possessors of a few hard-won clinical diagnostic skills based on rudimentary anatomy and physiology. In large measure we were observers. We lacked the drugs and procedures with which to intervene. By the turn of the century, motivated in part by the Flexner Report, medicine began to become more "scientific." The emphasis became rigorous science rather than art. The anecdote and series were displaced by the clinical trial. Medical research was to be based on the "scientific method," but little thought was afforded to its limitations.

The scientific method is predicated on some stringent preconditions. For it to succeed as an analytic tool all factors materially affecting the outcome must be identifiable, controllable, and measurable. Only when all parameters but one, the independent treatment variable, can be frozen, may the effect of our manipulation of the independent variable on the disease process be measured. Ideally, subjects should be homogeneous, especially insofar as parameters which may influence outcome are concerned. Of particular concern in biology is the experimental situation in which an influential parameter is not

[1]Supported in part by Health and Welfare Canada and the National Cancer Institute, Canada. H. Schipper is a Clinical Trials Scholar of the National Cancer Institute of Canada.

[2]Manitoba Cancer Treatment and Research Foundation, Winnipeg, Canada.

* *Reprint requests to:* H. Schipper, Hematology/Oncology Service, St. Boniface General Hospital, University of Manitoba, Winnipeg, Manitoba, Canada R2H 2A6.

recognized. In this situation, for example, one arm of an experimental drug trial may appear to demonstrate a superior outcome, when in fact the difference observed represents the influence of the unrecognized parameter which was inequally distributed across treatment arms. Numerous lung cancer studies have encountered this difficulty. Further, the scientific method gains power when experiments are repeatable. If the study material is not uniform, at worst the experiment will not be repeatable and at best the result will lack generalizability.

The scientific method works in the basic sciences. We can isolate processes in test tubes or cyclotrons. We can identify parameters (eg, temperature, color, and solubility) and measure them (eg, degrees centigrade, spectral frequency, and milligrams per milliliter). We can also achieve experimental control (eg, by water baths, filters, and scales). The net result is repeatable science, having accurate, precise outcomes that can be interpreted and generalized. Moreover, results can be viewed dispassionately in the light of "cold science." However, when one moves from the physics or chemistry laboratory to the study of intact organs or animal systems, it becomes much more difficult. Confounding parameters are infinitely more numerous, harder to identify, and much more difficult to control. [For example, the serendipitous identification of cisplatin resulted from failure to consider the influence of electrode composition on a microbiology experiment (2).] The effect of this is more variability in experimental outcome and greater statistical variance.

Clinical trials mount an almost impossible challenge to the scientific method. They represent the ultimate step in the medical research chain of events leading (not necessarily sequentially) from laboratory science to clinical practice. The subject matter (patients) defies every defining characteristic of the scientific method. Despite all of our efforts at confirming pathology and establishing rigorous staging procedures, patients are far from homogeneous. Those factors which affect human social and biologic functions are not easily identified and are likely interactive. They certainly cannot be controlled.

Despite these apparent shortcomings, the clinical trials process has worked. Often cited is the evolution of treatment for childhood leukemia and Hodgkin's disease, where fatal illnesses have been rendered curable. However, it equally can be argued that beyond defining a standard of therapy and monitoring, the trials process contributed little that the immense power of the drugs would not have made apparent even without benefit of the randomized trial. (The same cisplatin never required a randomized trial to establish its role as the dominant therapy for testicular carcinoma.) Viewed from another perspective, the sequential trials demonstrating the therapeutic equivalence of lesser surgical procedures for primary breast cancer provide an important example of a clinical trials process being used to alleviate toxicity rather than alter the course of disease (3,4).

Where therapy has a marginal benefit, the ultimate shortcoming of the scientific clinical trial becomes manifest. Scientists are supposed to be dispassionate observers and the subjects of the experiment inherently neutral (ie, guinea pigs), their behavior not a function of endpoint selection. But patients are passionately involved in the scientific process and their clinical trials behavior is profoundly influenced by a set of endpoints which are deeply personal and quite different from those held important by clinical scientists. Clinical trials analysts measure tumor size, disappearance, reappearance, and survival. Patients measure quality of life. What is important to the scientist, the very numbers by which he judges the success of his therapy and plans the next steps, may not be relevant to patients.

We have not used quality of life as a clinical trials outcome measure mostly because we don't consider it scientific. Quality of life is viewed as an undefinable and unmeasurable gestalt. Measures of survival, response rates, etc, are held to be "hard," meaning that they are repeatable and can be represented by numbers expressed to several decimal places. Much evidence suggests the fallacy of this belief. So-called hard data relating to tumor size (5), response rates, wbc counts, and even survival have been found wanting, particularly at critical decision points, such as a wbc count of 4000/mm^3 which serves as the drug attenuation cut-off for many cancer therapies. On the other hand, quality of life is definable and measurable. Moreover, as an end of process marker, the endpoint relevant to the patient, it holds the promise of being the most sensitive and powerful measure of all (6).

Strategies for Developing Quality of Life Measures

The measurement of quality of life is an emerging science. It is also, at the moment, very fashionable. That is a bad combination. In our headlong rush to activate an important and popular concept, it is essential to pace research and development so that scientific rigor is not compromised. To proceed otherwise is to risk bad research and faulty data and to diminish the credibility of the effort.

The most difficult phase of evaluating quality of life is defining what is to be measured. The present state of the art mandates a pragmatic, function-oriented approach. We lack the tools to quantitate philosophic issues. Moreover, we would likely find it impossible to derive enough philosophic-demographic homogeneity to make meaningful comparisons. Similarly, it is unlikely that quality of life can be measured against some external standard. One man's success is often another's failure. However, by using subjects as their own controls one can circumvent the problems of patient inhomogeneity and absolute standardization in one step. Most important, one must understand that quality of life is a continuous variable, an ongoing response to events affecting the patient. Of prime

significance is not a patient's measured quality of life at a point in time, but rather the pattern of function/dysfunction over time. Using such a model, the functional impact of quality of life can be integrated longitudinally and comparisons between individuals or groups undertaken.

Thus, an acute illness (such as a pulmonary embolism) may show one pattern of dysfunction (fig 1), a chronic progressive illness another (fig 2), and a chronic disease with exacerbations and remissions a third (fig 3).

What emerges as a strategy is a functional definition of quality of life, measurable and evaluable over time. The patient serves as his own internal control, though the issues to be included in the assessment should reflect the concerns of family, patient, and health care providers.

Despite what may appear to be a certain conceptual vagueness, four central components have been consistently identified as substantive elements of quality of life: physical/occupational function; psychological state; sociability; and somatic discomfort (7-9). In certain disease states, other elements take on increased emphasis: sexuality in breast cancer and diseases involving the sexual organs (10,11) and speech dysfunction in laryngeal cancer (12). However, the generalizability of the four dominant factors offers substantial credibility to the evaluation process.

There are other strategies which seek to model human behavior as a series of decision points. In this form of analysis, subjects who are not necessarily patients are confronted with hypothetical or actual situations, and patients are asked to choose between therapies which offer qualitatively different outcomes. Variations of this "decision analysis" theme include the Q sort (13). A telling observation which derives from studies employing these techniques is that patients frequently prefer function-preserving treatments to more radical curative attempts,

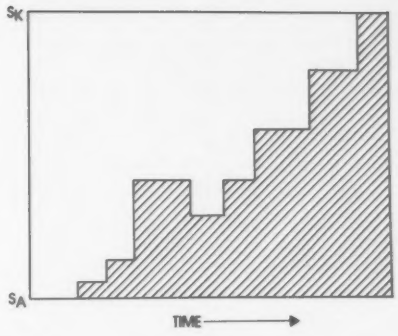


FIGURE 2.—Chronic progressive illness. In comparing outcomes, it is essential to consider both cumulative dysfunction and the pattern of such dysfunction. We may be confronted by outcomes of equivalent total dysfunction over time, but manifest in markedly different ways.

even at the expense of survival duration. This holds true not only in the palliative situation, but also where the primary intent is curative. These results suggest a clear pathway by which quality of life outcome measures may substantively influence clinical trials. Improvements to the survival-inferior but function-superior arms of a cooperative trial may be more productive than progressive augmentation of more toxic therapies. Furthermore, such approaches may ease the problem clinical trials designers face when attempting to "detoxify" successful treatments by withdrawing drugs or attenuating schedules.

Function should not be interpreted as a single variable. Inasmuch as quality of life is a composite of factors in a given disease or in a given individual, certain elements of

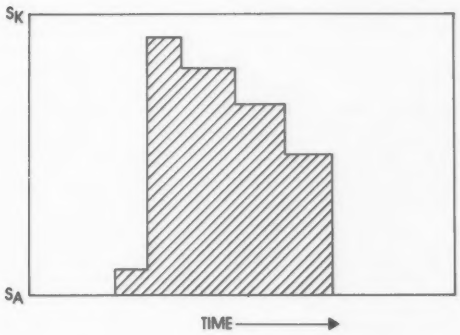


FIGURE 1.—Acute illness. Total dysfunction is represented by the shaded area under the curve. S_A and S_K represent defined levels or states of function.

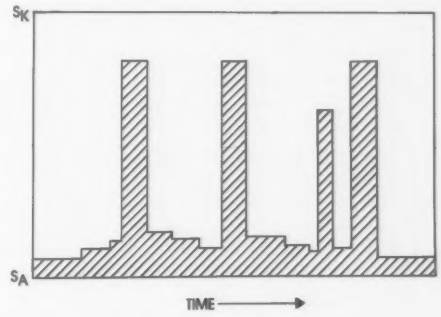


FIGURE 3.—Chronic disease with exacerbations and remissions. To some extent such a pattern is seen in patients receiving chemotherapy. However, there is evidence to suggest that many patients handle apparently equivalent morbidities associated with disease and treatment in profoundly different ways.

global quality of life may be differentially affected, so that it is possible to make necessary functional choice decisions. Thus, for example, Sugarbaker et al (14) found that patients undergoing radical limb-sparing procedures for sarcoma preserved a component of limb function at the expense of significant sexual compromise. If we develop convenient, reliable tools to make these observations, we will be able to adjust our therapies appropriately.

Methods and Pitfalls

The territory is new. While broad health status measurement is largely an established science, the use of measures such as the Sickness Impact Profile and Barthel Index are limited by their length and complexity (15,16). Furthermore, indices derived from bases of normal patients may not adequately reflect the quality of life of patients already experiencing significant illness. Conversely, psychometric measures derived for patients with manifest psychiatric disorders likely should not be extrapolated to patients whose chronic medical illness is the central issue.

It will likely be necessary to compromise some detail and precision to achieve the convenience and short-term repeatability required to enable these tools to measure quality of life longitudinally. In making these compromises it will be important not to diminish the sensitivity of the assay. An approach to consider is the use of specific indices focusing on quality of life issues of particular interest in a given disease. In adapting this strategy, however, great caution will have to be exercised to avoid the development of a plethora of incompatible indices. This would make it impossible to evaluate results across the broad range of clinical trials.

Some basic properties and characteristics for a clinical trials-oriented quality of life index can be set out using cancer as an example:

(*a*) The index should be cancer-specific, that is, the measure is specific enough to the cancer population to detect differences in functional state among cancer patients of a given disease group. In other words, unlike broad-based medical quality of life indices that are designed to measure the medical/functional state of free-living populations, this test should take into account that patients have already been diagnosed as having malignant disease and should concentrate on distinguishing functional states within this population.

(*b*) The index should be functionally oriented, addressing itself to those day-to-day living issues that represent the global construct of functional quality of life.

(*c*) It should be designed for patient self-administration and not require the intervention of interviewers or health professionals for its administration.

(*d*) The questions designed should be of general applicability, ease and consistency of interpretation, and of a limited number, to permit high compliance despite repeated administration.

(*e*) It should be repeatable, so that the patient's score can be followed over a period of time to elicit and evaluate trends both within patients and between groups.

(*f*) It should be sensitive across the range of clinical practice, being able to distinguish not only patients who are obviously well from those terminally ill, but more significantly, degrees of dysfunction between patients with varying extents of disease and intensities of therapeutic intervention.

(*g*) The instrument designed should have adequately demonstrated face, content, construct, and concurrent validity as well as reliability. These technical concepts will be further discussed.

Quality of life measures will be useless unless rigorously validated and their reliability established. While the reader is referred to any of a number of excellent texts for a fuller discussion, a brief overview of some basic concepts may be of assistance (17–20).

Validation is an inferential process. Since one assumes in developing a new test no fully acceptable tool exists to measure the property at hand (or why design a new one?), there is no absolute standard for comparison. The investigator is therefore forced to deduce from tangential sources that the new tool measures what it is supposed to.

There are a series of validation concepts to be considered: face, construct, concurrent, and predictive validities.

Face validity is a measure of the common sense of an instrument. Do the tests selected, the questions asked, or the measurements taken seem logically related to the phenomenon being measured? Some years ago a financial analyst made the observation that the Dow-Jones Index over a period of time bore a curious but highly correlated relationship to the temperature in Bombay measured 2 days previously. Though the Bombay weather office may have been deluged by the curious, the meteorological data have not to my knowledge been incorporated into the Dow-Jones because it made no sense. It had no face validity. It is also important to note that face validity is profoundly influenced by the experimental designer's overall model of the process under investigation.

Face validity has broader implications than overt common sense. It also reflects the completeness with which the tool reflects the overall process being examined. Does the tool reflect all influences thought to be reflected in the characteristic to be measured. A common pitfall in quality of life studies is to consider only physician- or patient-perceived factors, rather than encompassing the global universe of concern. If face validity is not properly considered, the tool will be found to correlate poorly with

the independent, parallel assessments which form the basis of face validity testing.

Construct validity is less directly explained. It is a measure of two distinct but closely related properties of any test which purports to measure the net outcome of several influences. The first property is the comprehensiveness of the instrument. The second property is the extent to which those factors identified and thought to influence overall outcome are the real influencing variables. In an idealized situation, the number of factors which influence the overall outcome are finite, identifiable, and precisely measurable. In this case, when the overall test is administered and the results are analyzed, there should be no unexplained variability in test results. To the extent that such variability exists, either there are other influences on overall outcome than those selected, or the test items chosen to represent a particular influence are inadequate to the task. In the type of psychometric testing used in quality of life studies, at least 75% of test variability should be accounted for by identifiable factors.

An example may offer some assistance. Suppose we were to design a nausea index. Possibly, we would consider the overall impact of nausea to be a composite of duration, intensity, and the amount of associated vomiting. We might therefore ask five questions or make five measurements which we believe to be relatively specific to each component factor. As an estimate of the duration we might ask whether the nausea lasted minutes, hours, or days. We might inquire as to how many times the subject vomited. As designers we may believe that the questions or tests we have selected not only address all the relevant contributory factors, but also that individual test items are specific to one or another contributory factor. However, it may be that the response to a given question is in fact influenced by more than one of the contributing factors, or of larger concern, reflects in large measure a previously unidentified or ignored contributor. Using statistical processes such as factor analysis, one can assess construct validity. Even if individual items have overlapping specificities, the overall questionnaire may be valid if it can be shown that the factors which emerge are those intended. The stability of the factor analysis when repeated using different populations is evidence for both the construct validity and the generalizability of the instrument. Failure to perform this type of analysis may result in a test which measures something quite unintended and is therefore uninterpretable.

Concurrent validity is tested by correlating the results of previously established and accepted measures against the new test. The challenge facing the experimenter is to select as many independent tests *and observers* as possible, while maintaining a manageable test evaluation package. Numbers of subjects must be sufficient to permit the multiple comparisons necessary, and the range of subjects on whom the test is validated must be representative of those for whom it is ultimately intended. If new test and concurrent validation measures have different clinical ranges of sensitivity which are not taken into account, interpretation of concurrent validation data will be confounded. Properly constructed concurrent validation procedures can provide reassurance of the sensitivity of the test.

The foregoing validation procedures can be performed on a cross-sectional patient sample, thus avoiding the costs and complexity of a longitudinal follow-up study. However, if predictive validity is considered to be a desirable attribute, then a group of patients must be followed over time to determine under what conditions a test result derived at one point can predict future events.

If a test is to be of value it must also meet certain reliability criteria. A reliable test is one which consistently scores the same value when measuring what is thought to be the same value of the process. For example, a reliable blood sugar measure should provide very similar results when tested on multiple occasions against a fixed solution of known concentration. The situation, however, is rendered more complex where human functional measures are concerned. In the first instance there may not be a measurable "gold standard" which can be held constant. Second, procedures have to be used to minimize or correct for the effect that repeat exposure to the test has on subject responses. The use of analog (Likert) scales (fig 4), randomizing question sequence and reversing the polarity of questions, can be helpful.

Patients usually want to please their health providers. This understandable tendency is a major, though often incompletely considered, factor which can fatally bias an otherwise useful test. "Social desirability" is a quasistatistical measure which can identify and suggest appropriate measures for correcting this natural bias (21).

Results to Date

The major advance in quality of life research in clinical trials is the growing realization that it is important. The issue is now widely discussed in the literature, and methodologies are beginning to evolve. An overview of the cancer literature provides insight into the nature of the development process and suggests future direction.

There is a large descriptive literature examining function (11,22,23), psychosocial state (24-26), and somatic sensation, notably pain and nausea and vomiting (27) in cancer patients. Sexual dysfunction and rehabilitation

FIGURE 4.—Analog (Likert) scale. Such scales are straight lines, frequently 10 cm in length. Ends of the line indicate polar opposites of opinion or outcome. Anchor points at midline or at intervals may be added. Considerable psychometric experience has established the sensitivity and reliability of this type of scale.

are also widely considered (28). A consensus has emerged, namely, that quality of life can be functionally described as a composite of four dominant contributors: physical/occupational function, psychological state, sociability, and somatic discomfort. What is as yet unclear is the relative importance of each component to overall quality of life. Even allowing for the influence primary question selection has on overall weightings which emerge from the factor analysis process, there is a remarkable consistency to be observed in the ordering of component weights across most published studies. Thus, it would appear to be an underlying process which is amenable to measurement.

The most established functional outcome measure employed in oncology is the Karnofsky scale (29). This scale and the compressed Eastern Cooperative Oncology Group (Zubrod) scale have been widely used to stratify patients for trial entry and analysis. It is clear that they are reasonable measures of physical function, but our data suggest that they are incomplete measures of overall functional quality of life.

The modern era of quality of life studies begins in 1976 when Priestman and Baum (30) proposed a ten-item quality of life questionnaire (for use in breast cancer patients whose sum score of 100 seemed to correlate with chemotherapy toxicity). In addition, when readministered at 3 months, the test seemed to be able to distinguish between those patients who had had an objective tumor response and those who had not. But the test was not adequately validated, the sample size was small (20 people), and the disease was narrowly defined. However, despite these shortcomings, their experience clearly demonstrated the viability of the concept. Later, Spitzer et al (31) took the process a step further and devised a five-question index which was designed to be physician- or nurse-administered. Subsequent attempts increased the rigor of the design and began to emphasize patient self-administration and broader reliability and validity testing, employing multiple observer and population techniques. The Functional Living Index for Cancer (FLIC) was developed at the Manitoba Cancer Treatment and Research Foundation (7). Development and validation subjects were 837 unselected patients representing the range of clinical practice, both inpatient and outpatient, encountered at two major Canadian cancer centers. The 22-item questionnaire was designed for repeated patient self-administration at frequent intervals (fig 5). Validation studies demonstrated sensitivity adequate to distinguish follow-up patients from those hospitalized and those hospitalized from those in palliative care units. Factor analyses on validation samples of 175, one from each cooperating institution, were remarkably stable, adding further to the weight of evidence that quality of life is a measurable and evaluable construct. Component factors representing physical function and psychological state correlated highly with specific concurrent validation tests, further suggesting that the FLIC and subsequent measures may be used not only as global estimates of quality of life but also as screening tools in the identification of specific dysfunction (table 1). To this point, however, there are no published reports evaluating the FLIC's performance in longitudinal studies. Recently, Selby et al (32) reported development of a 31-item test for use in patients with breast cancer. Particular emphasis (12 of 31 test items) was placed on symptoms especially related to breast cancer and chemotherapy, but it is unclear whether the disease specificity of the index augmented its clinical sensitivity. Of greater importance, however, was a further demonstration, in a well-validated study, of the consistency of component factors making up the quality of life construct. These studies were all cross-sectionally validated (7-9,32).

Current studies employing the more cumbersome but ultimately essential longitudinal study approach are now underway, frequently in conjunction with major clinical trials groups. In the process, measured quality of life is being incorporated into cancer clinical trials as an outcome measure.

The Future

When fully established, quality of life outcome measures will have a powerful influence on the clinical trials process.

First, we will have an outcome measure which measures the end product of the health care system insofar as the patient is concerned. Our tool will provide insight and direction, useful in solving accrual and compliance problems, which currently retard reasonable clinical trials. If, as the Community Clinical Oncology Program data suggest (R. Frelick, personal communication), only 20% of eligible patients are entered in trials, the generalizability of our research is seriously compromised and our obligations to deliver health care are not being met.

More specifically, the application of quality of life indices will influence clinical research in palliative care and curative therapies in different ways. When it is not possible to offer a reasonable prospect for cure, longevity per se ought not be the *sine qua non*. Palliation by definition means preservation of quality of life. To a considerable extent the hospice movement has focused on pain; yet, clearly as our data and those of others make clear, somatic discomfort is only part of the total quality of life construct. The availability of quality of life measures will assist in refocusing toward these broader goals. Anticancer and quality of life management ought not to be considered mutually exclusive alternatives as is now too often the case.

Where intensive therapies are being evaluated, the addition of quality of life assessment to more conventional outcome measures may both provide explanations for apparent discrepancies between patient function and

MANITOBA CANCER TREATMENT & RESEARCH FOUNDATION
FUNCTIONAL LIVING INDEX: CANCER (FLIC)

Date_____

1. How well do you appear today?

1 — 2 — 3 — 4 — 5 — 6 — 7
Extremely Poor — Extremely Well

2. Rate your confidence in your prescribed course of treatment.

1 — 2 — 3 — 4 — 5 — 6 — 7
Confidence — Very Confident

3. How much of your pain or discomfort over the past 2 weeks was related to your cancer?

1 — 2 — 3 — 4 — 5 — 6 — 7
None — All

4. Rate how willing you were to see and spend time with friends, in the past 2 weeks.

1 — 2 — 3 — 4 — 5 — 6 — 7
Unwilling — Very Willing

5. Rate the degree to which you are frightened of the future.

1 — 2 — 3 — 4 — 5 — 6 — 7
Constantly Terrified — Not Afraid

6. How much nausea have you had in the past 2 weeks.

1 — 2 — 3 — 4 — 5 — 6 — 7
None — A Great Deal

7. Rate how willing you were to see and spend time with those closest to you, in the past 2 weeks.

1 — 2 — 3 — 4 — 5 — 6 — 7
Unwilling — Very Willing

8. How much of your usual household tasks are you able to complete?

1 — 2 — 3 — 4 — 5 — 6 — 7
All — None

9. Rate the degree to which your cancer has imposed a hardship on you (personally) in the past 2 weeks.

1 — 2 — 3 — 4 — 5 — 6 — 7
Tremendous Hardship — No Hardship

10. How much is pain or discomfort interfering with your daily activities?

1 — 2 — 3 — 4 — 5 — 6 — 7
Not At All — A Great Deal

11. Rate in your opinion, how disruptive your cancer has been to those closest to you in the past 2 weeks

1 — 2 — 3 — 4 — 5 — 6 — 7
Totally Disruptive — No Disruption

TEAM NUMBER **A** **168** **№** **5**

FIGURE 5.—Example of questions from FLIC. Several principles of questionnaire design
tions varies. (*c*) Questions are not arranged in subject groups. (*d*) Overlapping items are

12. How uncomfortable do you feel today?

1 — 2 — 3 — 4 — 5 — 6 — 7

Not At All — Very Uncomfortable

13. Rate your satisfaction with your work and your jobs around the house in the past month.

1 — 2 — 3 — 4 — 5 — 6 — 7

Very Dissatisfied — Very Satisfied

14. Rate how often you feel discouraged about your life.

1 — 2 — 3 — 4 — 5 — 6 — 7

...ways — Never

15. Rate the degree to which your cancer has imposed a hardship on those closest to you in the past 2 weeks

1 — 2 — 3 — 4 — 5 — 6 — 7

No Hardship — Tremendous Hardship

16. Do you feel well enough to make a meal or do minor household repairs today?

1 — 2 — 3 — 4 — 5 — 6 — 7

Very Able — Not Able

17. How well do you feel today?

1 — 2 — 3 — 4 — 5 — 6 — 7

Extremely Poor — Extremely Well

18. Has nausea affected your daily functioning?

1 — 2 — 3 — 4 — 5 — 6 — 7

Not At All — A Great Deal

19. Rate your ability to maintain your usual recreation or leisure activities.

1 — 2 — 3 — 4 — 5 — 6 — 7

Able — Unable

20. How much time do you spend thinking about your illness?

1 — 2 — 3 — 4 — 5 — 6 — 7

Constantly — Never

21. How well are you coping with your everyday stress?

1 — 2 — 3 — 4 — 5 — 6 — 7

Not Well — Very Well

22. Most people experience some feelings of depression at times. Rate how often you feel these feelings.

1 — 2 — 3 — 4 — 5 — 6 — 7

Never — Continually

PLEASE INDICATE WITH ✓ YOUR RATING

Please check to see if you have completed all questions. Then, fold on the perforated line, seal, and drop in the mail. Thank you for your valuable assistance in this project.

...ire design are illustrated. (*a*) The analog scales are of identical length. (*b*) The polarity of ques-...items are designed to increase validity of the test and amplify scalar sensitivity.

TABLE 1.—Correlations of factor scores with concurrent validation tests: comparison of Winnipeg (Run 2) and Edmonton data *

Validation test	Physical well-being and ability factor of FLIC		Emotional state factor of FLIC	
	Winnipeg (Run 3)	Edmonton	Winnipeg (Run 3)	Edmonton
Katz activities of daily living	-0.228	-0.207	0.042	0.20
General health questionnaire				
A scale: somatic symptoms	-0.541	-0.606	-0.315	-0.346
B scale: anxiety and insomnia	-0.204	-0.224	-0.557	-0.543
C scale: social dysfunction	-0.649	-0.571	-0.279	-0.083
D scale: severe depression	-0.263	-0.328	-0.556	-0.427
Total	-0.549	-0.572	-0.594	-0.469
Beck depression	-0.567	-0.640	-0.517	-0.528
Karnofsky	0.757	0.554	0.066	0.157
Spielberger				
State anxiety	-0.308	-0.326	-0.629	-0.588
Trait anxiety	-0.296	-0.338	-0.735	-0.609
Melzack				
Present pain index	-0.657	—	-0.067	—
Pain rating index	-0.555	—	-0.128	—

* Several principles of concurrent validity testing are illustrated by this example. (*a*) Each factor to be measured is correlated against a No. of established measures, using multiple independent observers. (*b*) Validation experiment is repeated using different groups (in this case Edmonton and Winnipeg), and correlations remain constant. (*c*) Correlations are appropriate, ie, the physical function parameter correlated well with other physical function measures ($r > \pm .54$), but poorly with psychological function measures ($r > \pm .32$). *From* ref 7 with permission from publisher.

clinical response and pose some new and potentially difficult but important questions for clinical trials investigators. For example, if, based on traditional criteria, two clinical trials provide similar outcomes, a significant discrepancy between quality of life outcomes would readily influence subsequent treatment selection. More perplexing, however, is the potential dilemma of divergent results, that is, where an outcome that is viewed inferior by classical criteria is associated with a significantly improved overall quality of life response. Which treatment arm does the investigator then recommend, the one with the greater response and/or survival or the one with a quality of life outcome that appears superior? The availability of a valid measure of quality of life provides alternatives. (*a*) It may be possible to alter therapy in the "traditionally" superior arm to improve the quality of life outcome. (*b*) If the former is not possible, at least we are in the position to offer patients not only response, disease-free survival, overall survival, and medical toxicity information, but also, in a reasonably concrete form, quality of life information. In such circumstances the patient will be better able to participate in making an informed decision.

To the extent we are unable to prevent disease, the ultimate goal of medicine is to effect a cure and return the patient to a life-style not diminished by the illness and its treatment. The scientific power of modern medicine provides the biologic rationale to intervene with a force never before possible. The rigorous development and application of quality of life assessment is neither unscientific nor in competition with basic medical science as we know it. It merely serves to emphasize that the proper study of mankind is man.

REFERENCES

1. SIMON RM. Design and conduct of clinical trials in cancer. *In* Cancer: Principles and Practice of Oncology (DeVita V, Hellman S, and Rosenberg SA, eds). Philadelphia, Lippincott, 1982, pp 198-225.
2. ROSENBERG B. Cisplatin: its history and possible mechanisms of action. *In* Cisplatin, Current Status and New Developments. New York, Academic Press, 1980, pp 9-20.
3. VERONESI U, SACCOZZI R, DEL VECCHIO M, ET AL. Comparing radical mastectomy with quadrantectomy, axillary dissection and radiotherapy in patients with small cancers of the breast. N Eng J Med 305:5-11, 1981.
4. FISHER B, REDMOND C, and FISHER ER. Participating NSABP investigators. The contribution of recent NSABP clinical trials of primary breast cancer therapy to an understanding of tumor biology: an overview of findings. Cancer 46:1009-1025, 1980.
5. WARR D, MCKINNEY S, and TANNOCK I. Influence of measurement, error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. J Clin Oncol 2:1040-1046, 1984.
6. SCHIPPER HL. Why measure quality of life. Can Med Assoc J 128:1367-1369, 1983.
7. SCHIPPER H, CLINCH J, MCMURRAY A, ET AL. Measuring the quality of life of cancer patients: the functional living index—cancer: development and validation. J Clin Oncol 2:472-483, 1984.
8. GOUGH IR, FURNIVAL CM, SCHILDER L, ET AL. Assessment of the quality of life of patients with advanced cancer. Eur J Cancer Clin Oncol 19:1161-1165, 1983.
9. PADILLA GV, PRESANT CA, CRANE M, ET AL. Quality of life index

for patients with cancer. Res Nurs Health 6:117-126, 1983.
10. SCHAIN WS. Sexual problems of patients with cancer. *In* Cancer: Principles and Practice of Oncology (DeVita V, Hellman S, and Rosenberg SA, eds). Philadelphia, Lippincott, 1982, pp 278-290.
11. MEYEROWITZ BE. Psychosocial correlates of breast cancer and its treatments. Psychol Bull 87:108-131, 1980.
12. MCNEIL BJ, WEICHELBAUM R, and PAUKER SJ. Speech and survival: trade-offs between quality and quantity of life in laryngeal cancer. N Engl J Med 305:982-987, 1981.
13. STEPHENSON W. The Study of Behavior: Q-Technique and Its Methodology. Chicago, University of Chicago Press, 1953.
14. SUGARBAKER PH, BAROFSKY I, ROSENBERG SA, ET AL. Quality of life assessment of patients in extremity sarcoma clinical trials. Surgery 91:17-23, 1981.
15. BERGNER M, BOBBIT RA, POLLARD WE, ET AL. The sickness impact profile: validation of a health status measure. Med Care 14:57-67, 1976.
16. MAHONEY FL, and BARTHEL DW. Functional evaluation: the Barthel Index. Md State Med J:61-65, 1965.
17. NUNNALLY JC. Psychometric Theory, 2nd ed. New York, McGraw-Hill, 1978.
18. THORNDICKE RL. Applied Psychometrics. Boston, Houghton Mifflin, 1982.
19. CRONBACK LJ. Essentials of Psychological Testing, 3rd ed. New York, Harper and Row, 1970.
20. BROWN FG. Principles of Educational and Psychological Testing, 3rd ed. New York, Rinehart and Winston, 1983.
21. JACKSON DN. Dynamics of structured personality tests. Psychol Rev 78:229-248, 1971.
22. VAN DAM FSAM, LINSSEN CAG, and COUZIJN AL. Evaluating "quality of life" in cancer clinical trials. *In* Clinical Cancer Trials: Methods and Practice (Buyse M, Staguel M, and Sylvester R, eds). Oxford, Oxford University Press, 1983.
23. SILBERFARB PM, PHILIBERT D, and LEVINE PM. Psychosocial aspects of neoplastic disease II. Affective and cognitive effects of chemotherapy in cancer patients. Am J Psychiat 137:597-601, 1980.
24. HERSH SP. Psychologic aspects of patients with cancer. *In* Cancer: Principles and Practice of Oncology (DeVita V, Hellman S, and Rosenberg SA, eds). Philadelphia, Lippincott, 1982, pp 267-277.
25. GORZYNSKI G, MASSIE MJ, and HOLLAND J. Diagnosis and management of psychiatric syndromes in cancer patients. *In* Psychopharmacology in Clinical Practice (Derogatis L, ed). New York, Addison-Wesley Press, 1981.
26. GOTTSCHALK LA. Measurement of mood and affect in cancer patients. Cancer 53:2236-2240, 1984.
27. MORROW GR. The assessment of nausea and vomiting: past problems, current issues, and suggestions for future research. Cancer 53:2267-2277, 1984.
28. GREENBERG DB. The measurement of sexual dysfunction on cancer patients. Cancer 53:2281-2284, 1984.
29. KARNOFSKY DA, and BURCHENAL JH. The clinical evaluation of chemotherapeutic agents in cancer. *In* Evaluation of Chemotherapeutic Agents (Macleod CM, ed). New York, Columbia University Press, 1949, pp 191-205.
30. PRIESTMAN TJ, and BAUM M. Evaluation of quality of life in patients receiving treatment for advanced breast cancer. Lancet 1:899-900, 1976.
31. SPITZER WO, DOBSON AJ, HALL J, ET AL. Measuring the quality of life of cancer patients. A concise QL-index for use by physicians. J Chronic Dis 34:585-598, 1981.
32. SELBY PJ, CHAPMAN JAW, ETAZADI-AMOLI J, ET AL. The development of a method for assessing the quality of life of cancer patients. Br J Cancer 50:13-22, 1984.

DISCUSSION

Q: (Dr. Arnie Weiss, Lederle) We in the pharmaceutical industry would like to be able to use something like your FLIC to actually differentiate between treatments. Do you have any evidence that, in fact, quality of life measures have ever done so successfully?

A: (Dr. Schipper) As far as whether quality of life issues have influenced trials and outcomes, we've got some preliminary data suggesting in fact that they do. Only now is the index starting to be used in a number of centers where we can accumulate those data. What we're really anxious to do is let them grow in the real world and make them better. We are using them as a stimulus for both discussion and further investigation in the area.

Q: You mentioned that you take repeat measurements on the same patients over time, and I understand that this causes difficult statistical problems. Do you have a recommendation on how to come up with an overall answer to the question of whether there is a difference between the two treatment groups in a trial, without getting into overly complex statistical analyses?

A: What you're doing is sampling the data over time. That's an issue that's distinct from how frequently you actually analyze and make public the data you are collecting. There are specific methodological concerns that go into designing any test that's designed to be used on a repeated basis. For example, patients are very keen to please their physicians and if you devise an index in which you are insensitive to this concern you may get biased answers that are designed to please you. Also, as you run the test on a repeated basis, established and definable biases start to emerge that are hard to deal with. In designing the test, we tried to account for that and according to my statistical and psychological colleagues, the test is remarkably clean of that kind of concern. Then there is the issue of test-retest reliability. Analog scales help you to cope with that issue and allow you to use the index repeatedly without training people to interpret answers. There are other statistical techniques, like flipping the poles around on successive questionnaires and changing the order and so on. So, we think it can be integrated over time in the manner of the concept diagrams I showed earlier.

Influence of Measurement Error on Response Rates

David Warr,* Steven McKinney, and Ian Tannock[1,2]

In reports of chemotherapy trials the frequency of complete plus partial responses is usually the most prominently displayed result. One disturbing aspect about this figure, however, is its variability. For example, the response rates for carcinoma of the colon treated with 5-FU range from 8% to 85% (1). This variation is usually attributed to differences in patient characteristics, different schedules of drug administration, disagreement over what constitutes an evaluable patient, or problems with small numbers, ie, a sampling error. There is, however, both clinical and experimental evidence that these explanations are not sufficient.

Davis et al (2) used the results of the Clinical Oncology Group (COG) chemotherapy trial for metastatic breast cancer to illustrate how response rates may be affected by differences in evaluation criteria. A single data set was re-analyzed according to the COG, Eastern Cooperative Oncology Group, Southwest Oncology Group, or Southeastern Cancer Group assessment criteria. The response rates for evaluable patients ranged from 38% to 52%. Similar changes in results after altering the response criteria have been reported in chemotherapy trials for prostatic cancer (3) and malignant melanoma (4).

In 1976, Moertel and Hanley (5) published the results of a study on measurement error which utilized simulated tumor nodules (STN). Their conclusions were that the category of minimal response was unreliable and that even a conventional definition for partial response (PR) would lead to a 5%–10% response rate due to error alone. Intrigued by these results, we have carried out a study that was similar in design but also used clinically relevant lesions (lymphadenopathy, hepatomegaly, metastases on chest roentgenograms) and examined the effect of altering response criteria (6).

STUDY DESIGN

The examiners in our study consisted of eight physicians: four medical oncologists and four medical oncology residents. The STN were 16 spheres ranging in diameter from 1.9 to 6.5 cm that were placed in random order on a foam rubber mattress and covered by another thin (1.3-cm) layer of foam rubber to simulate skinfold thickness. Seventeen chest roentgenograms (posteroanterior view) showing a total of 37 clearly defined metastases were selected. Fourteen patients with head and neck cancer or lymphoma associated with enlarged neck nodes volunteered for assessment. Fifteen patients with hepatic enlargement due to metastatic cancer also agreed to participate.

The maximum and perpendicular maximum diameters of the lesions (excluding those in the liver) were recorded by each physician and then multiplied to give an estimate of cross-sectional area. For the liver, the span below the right costal margin was added to the span below the xiphisternum to give a single linear measurement. To calculate the false response rates due to measurement error, we then compared all possible pairs of area estimates for a single lesion. A false PR was defined as an apparent shrinkage of $> 50\%$. A false minimal response was defined as an apparent shrinkage of 25%–50%. A false disease progression was defined as an increase $\geq 25\%$. The effect on false PR rate of measuring more than one lesion was simulated by comparing the sum of the estimates of cross-sectional area for two and for three specified lesions. By comparing triplets or quadruplets of measurements on a single lesion, it was possible to simulate the effect of requiring that two or three successive measurements must show at least a 50% reduction prior to declaring a PR.

RESULTS

Figure 1 shows the relationship between lesion size and the probability of false PR. The error rates for the STN were clearly greater for smaller lesions, but a similar variation was not apparent for the neck nodes. Thus,

[1]Department of Medicine, Princess Margaret Hospital and Ontario Cancer Institute, Toronto, Canada.

[2]We thank the *Journal of Clinical Oncology* for permission to use material from a previously published article. Copyright 1984, American Society of Clinical Oncology.

* *Reprint requests to:* David Warr, MD, Department of Medicine, Princess Margaret Hospital and Ontario Cancer Institute, 500 Sherbourne St, Toronto, Ontario, Canada M4X 1K9.

there were other factors influencing the reliability of measurement.

The frequency of false PR, false PR plus minimal response, and false disease progression for all lesions is shown in table 1. Similar to the findings of Moertel and Hanley (5), the false PR rates for the liver size, neck nodes, and smaller STN were approximately 10%. The results for the lung metastases were much better, presumably because these lesions were immobile and flat in contradistinction to the other tumors. Inclusion of minimal response as an indicator of antitumor activity increased the error rate substantially (table 1). Similar results were obtained for the category of false disease progression. To lower the error rates to a level identical to those for PR would require redefining disease progression as a ⩾ 100% increase in area.

When two or three lesions were assessed (rather than a single index lesion), the false categorization frequency decreased (table 2). Changing the definition of PR to require that all lesions decrease in area by 50% rather than simply the sum of areas led to a further decline in PR as shown in table 2.

TABLE 1.—False categorization of response from comparison of all pairs of measurements on the same lesion

Measurement	% false categorization*		
	PR	PR + minimal response	Disease progression
Simulated nodules (1.0-2.6 cm)	12.6	31.0	34.3
Simulated nodules (3.2-6.5 cm)	1.3	19.7	24.0
Neck nodes	13.1	32.1	33.4
Lung metastases (chest roentgenograms)	0.8	11.2	15.9
Liver size†			
> 50 decrease	8.5		28.7
> 30% decrease	18.4		

*PR was a > 50% decrease in area, PR + minimal response was a > 25% decrease in area, and disease progression was a > 25% increase in area, except for the liver size.

†Decrease in the sum of the linear measurements of the liver edge below the costal margin in the midline and midclavicular line (MCL).

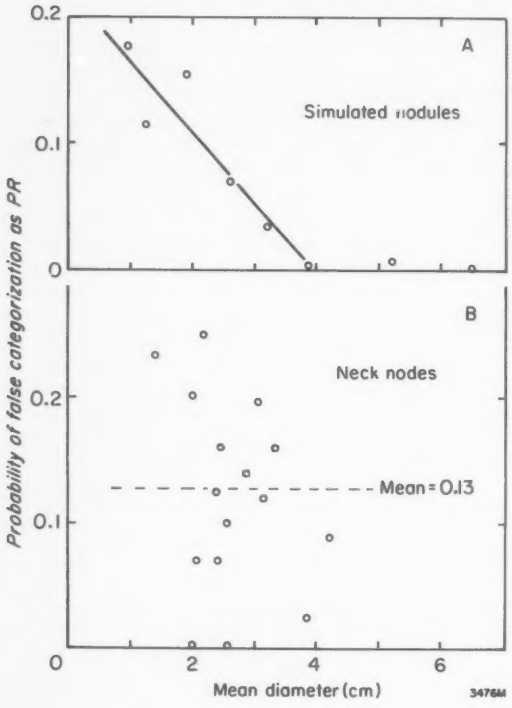


FIGURE 1.—Relationship between the probability of false categorization of PR and (A) diameter of the simulated nodule and (B) mean diameter of cervical nodes. Regression line was fitted to the data of 1A (r^2 = .86). There is no significant correlation in 1B (r^2 = .07).

The definition of PR may include a requirement that two or more successive measurements demonstrate a 50% reduction in area when compared to the initial size. Table 3 shows the effect on false PR of these more stringent criteria.

The underlying assumption in these results is that, with ineffective treatment, tumor size will not change between assessments. This assumption can be changed to allow for tumor growth. Figure 2 is a power curve derived from the lymph node data by comparing measurements on lesions of different size. It relates the relative change in cross-sectional area to the probability of declaring that change to be a PR. A typical volume doubling time for carcinomas is 3 months (7), which would translate into an increase in cross-sectional area at 1 month (the usual time of reassessment) of only 17%. From the power curve there would still be approximately a 7% chance of declaring a false PR. Similar calculations have been carried out for more rapidly or more slowly growing lesions in table 4.

DISCUSSION

Our results with clinically relevant lesions were similar to those of Moertel and Hanley (5) but still may not accurately portray those in clinical practice. All of the lesions had clearly palpable or visible margins, a much less common finding in the typical chemotherapy patient. The examiners were aware of the study and this may have led to more careful measurements. We have also not taken into account physician bias; our desire to achieve tumor shrinkage may cause subsequent measurements to underestimate the true size. All of these consid-

TABLE 2.—False categorization of PR (%) based on 1, 2, or 3 index lesions

Measurement	Single lesion	Response criteria* A: 2 lesions	A: 3 lesions	B: 2 lesions	B: 3 lesions
Simulated nodules (1.0–2.6 cm)	12.6	6.6	3.2	2.0	0.3
Simulated nodules (3.2–6.5 cm)	1.3	0.2	0.03	0.03	0
Neck nodes	13.1	5.0	2.4	2.1	0.4
Lung metastases	0.8	0.1	0.03	0.06	0

*Criterion A is the sum of areas decreased by > 50%, with no lesion increasing by > 25%. Criterion B is that all lesions are required to decrease by > 50%. Criterion A is in current clinical use; criterion B is more stringent.

erations would tend to increase the false response rates that we have calculated.

For phase III trials palliation and survival are the most relevant outcomes, yet the response rate is often regarded as the primary indicator of patient benefit. We believe that this practice will continue until more refined measures of palliation are developed. The results of this study plus our own clinical experience have been used to suggest criteria which will decrease the number of false responses and exclude short-lived shrinkage (which rarely translates into worthwhile symptomatic improvement). The definitions listed in table 5 cannot substitute for more appropriate endpoints, but may allow a more realistic assessment of the potential for patient benefit.

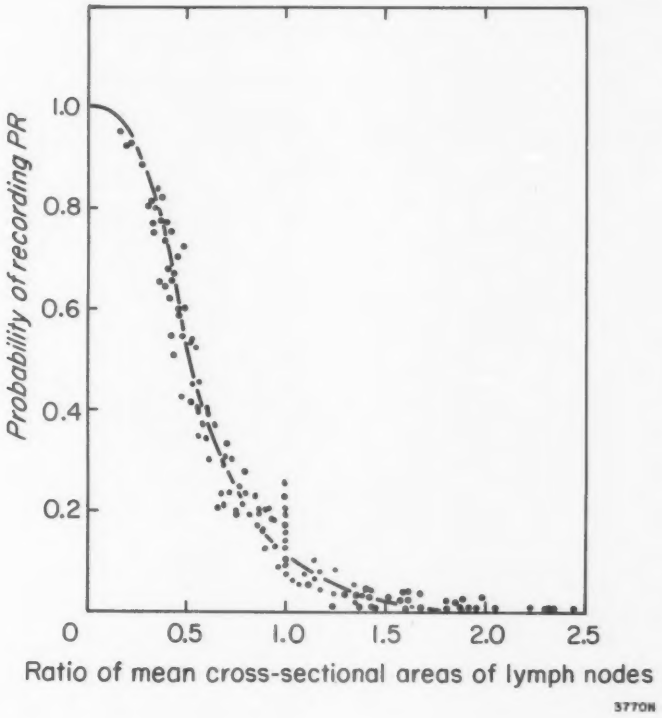


FIGURE 2.—Power curve illustrating the probability of recording PR for a given ratio of mean cross-sectional areas of cervical lymph nodes.

TABLE 3.—Influence of sequential measurements on false categorization of PR

No. of sequential measurements required for PR*	% false categorization of PR†	
	Lymph nodes	Liver
1‡	13.1	18.4
2§	4.6	5.0

*No. of consecutive measurements that must demonstrate the specified reduction in size to declare a PR.

†For lymph nodes, PR was defined as a > 50% decrease in area. For liver size, PR was defined as a > 30% decrease in the sum of the linear measurements of the liver edge below the costal margin in the midline and MCL.

‡Criteria used in many clinics.

§Criteria used by some cooperative groups.

TABLE 4.—Influence of rate of tumor growth on false categorization of PR for malignant neck nodes*

Volume doubling time of tumor (mos)	Change in cross-sectional area at 1 mo	% false categorization of PR
1.0	1.59	2
3.0	1.17	7
6.0	1.08	10
Infinite	1.00	13

*Nodes were re-evaluated 1 mo after initiation of ineffective treatment.

TABLE 5.—Proposed changes in criteria for tumor response: phase III trials

Criteria	Suggested changes
Disease progression	Redefine as a 100% increase in area* of any measurable lesion, or the appearance of new lesions
Minimal response	Term should be abandoned
PR	Redefine as a 50% decrease in area* of all index lesions recorded at 2 consecutive assessments at least 1 mo apart; no new lesions should occur; require a minimum duration† of 3 mos; require index lesions to have an initial mean diameter of 2 cm
Complete remission	Require disappearance of all evidence of disease recorded at 2 consecutive assessments at least 1 mo apart; require a minimum duration† of 3 mos

*For malignant hepatomegaly replace "area" with sum of measurements of the liver edge below the costal margin in the midline and MCL. Distance of the liver edge below the costal margin in the midline or MCL must exceed 5 cm to use hepatic disease as a marker of response.

†Duration of response is the interval between start of treatment and last assessment at which response criteria are satisfied.

REFERENCES

1. MOERTEL CG, THYNNE GS, HOLLAND JF, ET AL, eds. Cancer Medicine, 2nd ed. Philadelphia, Lea and Febiger, 1982.
2. DAVIS HL, JR, MULTHAUF P, and KLOTZ J. Comparisons of cooperative group evaluation criteria for multiple-drug therapy for breast cancer. Cancer Treat Rep 64:507-517, 1980.
3. YAGODA A, WATSON RC, NATALE RB, ET AL. A critical analysis of response criteria in patients with prostatic cancer treated with *cis*-diammine-dichloride platinum II. Cancer 44:1553-1562, 1979.
4. National Cancer Institute of Canada Melanoma Group. Vinblastine, bleomycin, and *cis*-platinum for the treatment of metastatic malignant melanoma. J Clin Oncol 2:131-134, 1984.
5. MOERTEL CG, and HANLEY JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. Cancer 38:388-394, 1976.
6. WARR D, MCKINNEY S, and TANNOCK I. Influence of measurement error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. J Clin Oncol 2:1040-1046, 1984.
7. STEEL GG, and LAMERTON LF. The growth rate of human tumours. Br J Cancer 20:74-86, 1966.

DISCUSSION

Q: (Dr. Susan Ellenberg, NCI) You've shown clearly that by making the criteria more stringent, you can lower the false-positive rate or the false-partial response rate. What you haven't said anything about is the false-negative rate. On what basis do you have confidence that by lowering the false-positive rate to the extent that you've done, you haven't raised the false-negative rate by an unacceptable amount?

A: (Dr. Warr) The evidence is actually contained in one of the illustrations from the article which is appearing in the *Journal of Clinical Oncology*. We did look at that because we are concerned about that. And I guess you do increase the false-negative rate somewhat—it isn't as dramatic as one might think. And you have to consider who the patients are whom you are then going to miss. They're the ones who have very borderline shrinkage that is very transient. If you feel that those sorts of responses are very important, then you've got an awful lot of problems. And the only way that you can really get around them is probably using about ten independent observers and taking an average of their results, which really isn't very practical.

Q: (Dr. Katia Tonkin, Princess Margaret Hospital) I'd just like to make a comment on a further study that has recently been completed, which is entitled "Criteria of Tumor Response in Cancer Chemotherapy Trials" [Ed. note: Tonkin et al, *Journal of Clinical Oncology*, June 1985]. In the phase III trials reviewed we found, not surprisingly, that the way criteria are actually reported is extremely variable and often incomplete. We were able to construct a hypothetical clinical trial where a response rate could vary from 75% to 20% depending on the stringency of criteria adopted. I agree with the consideration of having a 3-month minimum duration of response before the patient is coded as a responder. I think this would help a great deal. The other factor which is very important is the requirement for two assessments, 1 month apart, before coding a patient as a responder. I think this type of criterion would provide a more valid and stringent way of following the patients. Patients who respond for less than 3 months rarely experience a worthwhile clinical benefit. With stringent and hopefully more uniform criteria of tumor response one could, in the future, meaningfully compare the results of trials.

Q: In your simulation did you have a spot which had no lump to feel? This type of test would be of interest in helping us assess our criteria for complete remission.

A: No, we considered that, but we thought that would be a little underhanded, and we did want to keep the cooperation of the examiners throughout this experiment.

Q: (Dr. Charles Moertel, Mayo) I primarily rise to cheer you—certainly a kindred spirit—and congratulate you for a far more elegant study than the one that we had originally performed, and certainly, I think, far more helpful. My only problem, really, is with the criteria that you proposed for declaring objective responses. I would feel that your data have primarily demonstrated that we simply cannot use objective response, regardless of how we declare it, as a reliable index of therapeutic activity, that we may use objective response to form a hypothesis of possible therapeutic activity but not to establish it. Therapeutic activity can really be established only by more elegant trials involving randomized controls and all the other features we've talked about here today. What I'm concerned about is that, in the real world, we conduct phase II trials with the realistic expectation of identifying therapeutic activity of an agent in a given disease setting. The types of patients we choose for phase II trials are not necessarily at all representative of patients with disease overall. You get the nice round pulmonary metastasis that might represent 5% of the overall disease you eventually want to treat. But what you're looking for is some evidence of therapeutic activity. Frequently the patients you choose for phase II trials are not in the most ideal therapeutic setting, particularly in instances where, for ethical reasons, you have to

expose a patient to some type of primary treatment before you can bring him into a phase II setting. I would frankly look with glee upon a drug that, in breast cancer, would create a 50% response rate for 2 months in previous CMF failures. Your criteria, however, would reject such a drug. I would hate to toss out this drug because you don't have a 20% response rate at 3 months. A number of tumor settings present the same problem. So I feel your criteria are really far too rigid for the true purpose of a phase II trial, which is to seek therapeutic activity, granted that we are going to test this therapeutic activity in a more definitive trial later on.

A: I would agree with some of the comments. I do think one can relax some of these criteria for true phase II trials that are only looking for evidence of biological activity, but, at a minimum, I would standardize them. You can say that you are really going to compare across trials that have two arms, but that isn't reality, and as long as there are clinical trials, people are going to compare Cooperative Group A's results with Cooperative Group B's. I'm also concerned that data from phase II trials become the basis for conclusions and recommendations which could have been justified only by the results of a phase III trial.

Q: (Dr. Brigid Leventhal, Johns Hopkins) Two points. First, did you find that the physicians were educable? I think a lot of the problem that all of us have in doing physical exams is that we don't then instantly sacrifice the patient and measure the lesion we're palpating. Did you try having the same ten people moving the balls around and seeing if they got better after you lifted off the sheet? If so, that might be a useful teaching aid for all of us. The second question is, wouldn't you like to extend your observations to CAT scans, which I think a lot of us are using now? It doesn't help me much to know that a lesion can or cannot be measured on a chest x-ray if that's really not what I'm going to use.

A: Oh, I agree entirely with the suggestion about the CT scans, and there are one or two articles that are coming out about that. I think that's very important. As far as the first point, we actually did have the physicians measure these nodules on three occasions, and there really was not any systematic improvement in their results, but of course we didn't try any intervention, giving them feedback, about the fact, for example, that for most balls they were underestimating by about 4 mm for all the lesions.

Q: One, I commend you, in Dr. Moertel's enthusiasm, for bubble-bursting. I'm sure that you would agree that, before we launch any new criteria upon the oncological community, any criteria we recommend would have to be fairly carefully validated and checked for reliability in meaningful clinical contexts. I think that I agree with Dr. Moertel that the actual context of the study that you're doing, and perhaps the disease site, may have some influence on the criteria to be used. You may be very interested, clinically, in a very small activity in a disease where you have absolutely nothing, perhaps like metastatic melanoma, where in some other diseases, you may require rather more substantial evidence of activity before you actually are interested in applying that in a clinical situation. And finally, in the all too meager measurement literature, I think only about a couple of papers have appeared that have criticized this tendency to dichotomize or trichotomize the data as you measure them into some criteria of response; maybe we ought to be using the actual measurements, and using some sort of quantitative assessment as opposed to simply slotting into categories.

A: I think those points are fair. One of the things I'd like you to take away from this is that you really can jack up your response rates as high as you want, but you have to realize that, the more lenient you become, the more measurement error creeps in. Eventually that tumor is going to appear to respond, because of measurement error alone.

PANEL DISCUSSION: General Issues

Panelists: Drs. Charles Coltman, Thomas Davis, David DeMets, and Charles Moertel

Dr. Charles Coltman: I'm not an investigator into clinical trials methodology, nor am I a biostatistician. I've been in this sort of uncomfortable position several times recently, and I feel like a Presbyterian layman discussing Catholicism with John Paul II. However, I must say that I've been persuaded to a varying degree by these six presentations. If I were, in a Likert format, to quantify by drawing a hash mark along the line from totally persuaded to not persuaded at all, there's a clear separation in my scoring. I would rank the presentations by Drs. Simon, O'Fallon, Gail, and Warr toward the totally persuaded side of that scale, as an interested and learning observer in the clinical trials business. On the other hand, I might rank Dr. Taylor's presentation to the right or more toward the not-persuaded end of the scale, because I'm not sure that her data are generalizable to the community of clinical trials physicians. In fact, with deference to Dr. Patterson, I'm sure that he would agree that internists tend to be different from surgeons. While I thought that Dr. James Luce was going to present data to support the fact that chemotherapy trials have the same bias, when I confronted him over coffee, he had no such data, just anecdotes. So I would think that those same kinds of trials or studies need to be done with medical oncologists and other participants in clinical trials. Slightly to the left of that, I think, I would include Dr. Schipper's excellent presentation, even though I've taken to his methodology in scoring these presentations. I think it's probably because it's a new experience for me as a clinician. I simply wonder whether one's desire, at any time during the serious illness, to be with friends might not be related to one's premorbid desire to be with friends.

Dr. David DeMets: There was some discussion earlier about the issue, introduced by Dr. O'Fallon, of having a monitoring committee. I congratulate and commend the people who have put effort into doing that. I would only add that, in fact, this model has been used almost 20 years in what I call the NIH model. There's a long tradition of that data-monitoring model and it does work. I think it starts back as early as the early days of the Coronary Drug Project. The issue of blinding the principal investigator was raised. Dr. Tom Chalmers has written several articles about the rationale for doing that, so I won't elaborate. Finally, Dr. O'Fallon's talk suggests that the whole issue of decisions to terminate early or to continue is a very complex one. About a year ago, I was asked to give a paper at a symposium similar to this. I reviewed the literature and put together a series of examples about the complexity in making a decision to terminate early or not to terminate early. I would agree that it takes more than a single statistician or a single clinician to make such a decision. You have to have teamwork.

My second major point is, and I guess I happen to be among those who would probably be viewed as a hardliner, that you should analyze what you randomize; once you deviate from that, you're in for all kinds of trouble. I once, with a colleague, did a survey of the *New England Journal of Medicine* clinical trials and found that, in 37 of 103 consecutive clinical trials, not all of the patients were analyzed who, in fact, were randomized. So it's a fairly popular technique to take patients out of the analysis.

Although Dr. Gail left the door open a little with respect to eligibility, even that has its problems. There's a classic example, at least I consider it to be a classic example, in the literature of the Anturane Re-infarction Trial. Here, approximately 60 patients were taken out because they were deemed to be ineligible, even though they had, in fact, been randomized. However, it turned out that among those patients who were declared ineligible, ten in the treatment group died, versus four in the control group. The omission of these

deaths had a substantial impact on the overall results of the trial. So something even as simple or as innocent as initial eligibility criteria is not so simple.

The issue of taking patients out because of adherence is a tough one; it has been debated, at least in the cardiology literature. You can get examples that are ridiculous, such as placebo adherers doing better than placebo nonadherers. The bottom line is that we aren't even smart enough to figure out which way the biases go. Biases can work for the therapy, or they can work against the therapy, and you can't figure it out.

The issue of including or excluding centers, I think, is a dangerous one, and I'm glad that Dr. Gail took the position he did. You can take a negative study and make it positive by eliminating some center, and you can take a positive study and make it negative by taking out some center. So I tnink you have to present all the data, present the experiment that you did, not the one that you wanted to do or had in mind.

Finally, the issue of dilution has been raised again with respect to several of the papers and comments from the audience. I think the dilution effect is always there because of patient noncompliance. It's not unique to the Zelen design at all. It's always there if you take the principle that I espoused, ie, that you analyze what you randomize. Then you have to take into account that dilution effect in the sample size calculation in the first place. Although Dr. Simon didn't talk about that, I would at least comment that there is an important issue that ought to be considered, and there is a fair amount of literature of how to adjust the sample size for noncompliance.

I'm not a social scientist and don't know much about the area with respect to quality of life, but I have been put into the position, several times, of having to design or be involved in clinical trials that had aspects of quality of life. I would agree with Dr. Schipper that we need to develop better instruments. The ones that I used had some serious problems. They were too long, and they were not particularly sensitive to subtle changes; they were very good at diagnosing disease (that is, whether patients were healthy or not healthy). However, in comparing treatment A versus treatment B, these instruments, at least many of them, were not very sensitive to the subtle changes that you may be interested in.

And finally, there was some discussion about whether the problem that Dr. Taylor commented on is unique to that particular surgery trial. I would only point out to you the Coronary Artery Surgery Study, a study involving by-pass surgery versus medical care. There were 900 patients randomized to that trial, and about 27,000 patients entered into a registry. Why? Because many centers refused even to randomize patients, but they were willing to participate in the registry. So the strength of the emotional feelings is not unique to any particular trial that she was involved with.

Dr. Charles Moertel: The issues raised with regard to patient exclusion and with regard to stopping time can certainly help us as we move ahead into what we hope are going to be better clinical trial designs. I'd like to raise the question of what we do with the clinical trials that we already have on our hands, particularly where we know that the ways in which the trial was conducted are flawed according to what I've heard presented this morning. For example, what do we do with a surgical adjuvant trial that was projected to take on 500 entries but in fact took on only 250 because there was repeated display of data and the study was stopped when the *P* value reached the .05 level? How do we handle it? Should this study be reported at all? Obviously, we've greatly increased the possibility that just chance alone allowed the survival curves to spread, whereas if we had entered the full 500 patients they may have come back together again.

I'd like to ask the panelists how we interpret such studies that have been stopped prematurely? How do we report them? What do the *P* values mean?

Dr. DeMets: I think it's just a mathematical fact that repeated testing affects the *P* values and you can't duck that issue. There are ways to be conservative in interpreting those *P* values. Even if you had not thought of it when you started the trial, at least some of these methods could help you in interpreting those *P* values. It's not as simple as looking at a *P* value as some people have already said. I take the view that these statistical methods are not rules (I don't like the word stopping rules) but are guidelines. We haven't been smart enough yet to figure out absolute decision rules. But as guidelines I think they can still be helpful.

Dr. Moertel: Along the same lines, I'd like to ask another question about the number of patients who are lost to a study. Certainly, as Dr. Simon has pointed out, a gain of 10% in cure rates for most malignant diseases would be a major step forward. And yet in some of our randomized studies, our surgical adjuvant trials, or others where cure is an objective, we find that 10%, 15%, 20%, 25% of the patients initially entered on a study are lost to analysis because of cancellation, ineligibility, loss to follow-up. Sometimes these cases are distributed between study arms in a biased manner. Where do we draw the line as to what magnitude of ineligibility and cancellation rates makes a study invalid? Is there a rule of thumb saying that if you are looking for a given difference and if you lost so many patients to ineligibilities and cancellations, the result is not valid?

Dr. DeMets: I don't know of any rule of thumb, but it's obviously a problem. While I share Dr. Schipper's view that we have to measure more than body counts, I still rely on the total survival status as a very primary endpoint; at least there I can have a chance of finding out the outcome of the patient. It seems to me that if you start a patient on a therapy, even if he didn't get the full course, so to speak, you still need to follow up on that patient. When it comes to partial regression or complete regression, you can't fill in those missing data. If the missing data could alter the final outcome, it seems to me you're in trouble.

Dr. Moertel: Suppose he never started? He got assigned to an arm and he said, "I don't like that one," or his physician didn't like it, and he cancelled. These are pure bias-producing problems; how much of that do we allow before we decide a study is not valid?

Dr. DeMets: I don't know how much you allow but if you're getting substantial percentages you have a design problem. The only way to fix this is to avoid it by not randomizing that kind of patient. In the trials that I have been involved with, we have spent a great deal of time trying to figure out how to enter into the trials the patients who we felt would be good compliers. There is no simple solution, but we tried to exclude certain kinds of patients, such as alcoholics, those with a chemical dependence, or those who live a long distance away. Second, you ought to delay the randomization until the patient and the physician are actually convinced they are going to do it. Even having done these things, you will still get people who won't comply. I don't know of any rule that says, for example, that after 20% cancellation the study is not good and should be thrown out. I wish I did.

Dr. Judith O'Fallon: I think possibly that if we made a real effort to explain exactly what goes on in our studies when we write them up for publication and if the journals will permit us to publish these statements and not say that this is extraneous information making the article too long or whatever, maybe we would be motivated to run our studies more carefully. If we really have to admit all the flaws and, more importantly, if our readers understand all the flaws that are in the studies, they will have a better idea of how to interpret the results and we will be better motivated to keep ourselves really clean.

Dr. Richard Simon: My feeling is that if we are talking about 10% or 15% differences in the long-term survival rate, and if we are talking about 15% exclusions of patients for reasons other than ineligibility, I don't think we can be comfortable that our conclusions are correct. So I think really to have reliable conclusions we need to be including all randomized eligible patients and to be using end points like survival, where we can track down the patients, whether they complied with the treatment or not, to find out whether or not they are alive.

Dr. Moertel: You would not include the ineligible patient even though he was randomized, is that correct?

Dr. Simon: My personal feeling is that as long as it's a uniform review of eligibility—in other words, you don't review eligibility more for some patients than for others because of what you happen to notice about them later—there are many situations where there is not a bias resulting from excluding patients due to ineligibility. But anything after that I am not comfortable with excluding.

Dr. Coltman: I would like to ask a question of Dr. O'Fallon and Dr. Moertel. As chairman of the Southwest Oncology Group (SWOG) and as the beneficiary of the Mayo Clinic recently by having Drs. Fleming and Green join our new statistical center in Seattle, I'm persuaded, and I think many in the leadership of our group are persuaded, that the argu-

ments for keeping the covers on studies are very critical to the success of our studies and proper analyses. Since we are now in a transition phase where we've done repeated analyses and where we now have common-knowledge information about ongoing studies, how would you handle that in a group? Would you simply continue to analyze those in the open or would you suddenly make everything under cover with respect to phase III trials?

Dr. O'Fallon: My preference would be to try to get an agreement among all the members that the policy henceforth will be to keep preliminary results under wraps, and then put everything under wraps.

Dr. Moertel: I'd certainly agree with that. I've seen so many problems develop in clinical trials where people guess what the results will be either because the survival curves on preliminary analyses are lying on top of one another or because on preliminary analyses they are spreading. An otherwise very good study may be left in a terrible dilemma because people stop entering cases and you never get to your agreed accrual. Then you have an insoluble problem of trying to interpret results. Of course the other problem is that you may have many patients already entered on a toxic program and those survival curves aren't spreading. The investigators will stop treating the patients. Both of these actions of the investigators are perfectly natural things to do. The oncologist out in practice is going to say that "my conscience, my ethics demand that I do this" as soon as you show him suggestive preliminary results.

Dr. Coltman: Close it on everything that has been open in the past, as well as all future phase III trials?

Dr. Moertel: Sure.

Dr. Coltman: I am a little concerned about the effect of this policy on the level of interest at Group meetings.

Dr. O'Fallon: As soon as your study is done or you are ready for final analysis you have some very nice studies to report; you discuss the studies that are ready to be reported. If you've got one that you are very seriously considering closing, that is going to be presented to the membership for consideration of closing early, you are going to spend 30–60 minutes at each meeting talking about some of these problems.

Dr. Mitchell Gail: There has been an assertion that doing this will improve the accrual or prevent the accrual from trailing off. I wonder if there are any objective data because the question of interest is an important one. I also wonder if there are objective data according to complete divulgence of treatment groups, coded divulgence of separate treatment, or no divulgence whatsoever. I can imagine settings where accrual might go down because you're covering up the data.

Dr. Moertel: A lot of our North Central Cancer Treatment Group (NCCTG) trials, I think, have been very tough ones and they are being conducted in the community. We have, for example, a randomized trial in postmenopausal breast cancer that includes an untreated control arm. The same thing in colon cancer. These are toughies to sustain. I think *because* we don't display preliminary results these studies are going to be completed. Yet I have been in other circumstances where the results were displayed in studies with far fewer accrual requirements, and where these studies were never completed by groups that had far greater available patient resources than we have in the NCCTG. I feel people sustain interest in the blinded studies, at least as well, if not better, than if they are allowed to constantly sneak a peek at the results.

Dr. DeMets: I want to comment again on the history of this NIH model in noncancer trials. More or less, the objectives of recruitment have been obtained in all of these trials with the data totally blinded to the clinical investigators.

Dr. Thomas Davis: I'd like to make just one comment which relates to a common theme addressed by several of our speakers here today. That theme is the optimal size of clinical trials. No longer can we conduct a small trial at a single institution and expect to get a definitive answer which will stand up over the long term. We are wasting our resources when we proceed in that direction. Simple statistical considerations demand that we pool our resources to a much larger degree in order to answer the important questions regarding cancer treatment. As individual clinical investigators, we won't have the opportunity to be first author on as many manuscripts. However, this is a trivial price to pay for a

vastly improved cancer clinical trials program.

Dr. Gail: I would like to hear further discussion of blinded study results and monitoring committees for a group that's doing, let's say eight or ten studies at one time. What if the principal investigator plus the statisticians, the people who compose the principal investigation committee, were handed that responsibility for all eight or nine trials? Would that work?

Dr. Moertel: I'm sorry. Do you mean that you have eight or nine principal investigators plus all of your statisticians gathered together as one large monitoring committee?

Dr. Gail: One large committee and it's going to try to make a decision about nine trials, whether or not to continue.

Dr. Moertel: I think the rumors would start spreading because, at least in our group, that would mean every institution was represented and knew the results of every study. We have only ten members. We find it very easy to keep confidential a given study with a monitoring committee consisting of principal investigator, statistician, and myself as group chairman. But I think if you start spreading to eight to ten people on these committees, you are essentially telling the whole group what's happening.

Dr. Gail: So you require eight separate meetings, too?

Dr. Moertel: Yup.

Dr. DeMets: One comment on that, Dr. Gail, is that at least in the NIH model I referred to, most of the investigators or members of that team are not actually putting patients into that trial. That is, they are really independent of the trial. They have no axe to grind, presumably, but they have expertise that's relevant to what's going on.

Dr. Gail: So there is a major increment in the amount of meeting time at least?

Dr. DeMets: Well, if you have eight studies in one group, I suppose you could construct a committee that might be able to monitor perhaps not all eight but more than one at a time. So you could be more efficient. In fact, I participated in a study with two ongoing trials, and one data-monitoring committee did both.

Dr. O'Fallon: We have a different committee for each study. We need a certain amount of minimum expertise. An oncologist who is specializing in lung cancer isn't going to be the best person to be monitoring a study in gastrointestinal cancer. You don't have to be physically in the room to communicate effectively. We do a fair amount of conversing over the telephone.

Dr. Susan Ellenberg: One area that presents problems is the interest in presenting abstracts at professional meetings. We all know how much in the way of interim results is presented at these annual meetings of the many societies of which we are members. I wonder if any of you who are members of cooperative groups could say whether your groups have any kind of policy on presentation of interim data at these meetings or whether or not the professional organizations themselves ought to be monitoring this in consideration of abstracts to be presented.

Dr. Coltman: If we do as has been suggested in the SWOG, there will be no interim data to present. So our abstract rate will decrease.

Dr. Moertel: No it won't, it will just slow down for about 6 months, and you'll submit better abstracts later on.

Dr. Coltman: I look forward to that magnificent moment.

Dr. Davis: I would like to applaud Dr. Taylor for undertaking the study of physicians involved in cancer clinical trials. Although her presentation was basically an extended anecdote concerning one group of physicians participating in a single clinical trial, I don't think she needs to make any apology for that. Her anecdote should stimulate all of us to learn more about how physicians behave in the clinical trials environment. The Eastern Cooperative Oncology Group (ECOG) recently utilized some of the information which Dr. Taylor presented here today in conducting one of its own breast cancer trials. We were encouraged by her observation that the National Surgical Adjuvant Project for Breast and Bowel Cancers (NSABP) had been able to increase accrual by 400% while keeping the rate of noncompliance down to about 8%. We were struggling to complete a study of similar design involving a no-treatment control arm, and hoped to improve our accrual to this study by adopting the same informed consent mechanism employed by NSABP. Unfortu-

nately, it turned out that what was true for surgeons in the NSABP was not true for medical oncologists in ECOG. The accrual on our study increased by only 80%, whereas our failure to comply with the assigned treatment rose to 25%. If only we had known a bit more about the behavior of medical oncologists in our own cooperative groups, we might have avoided making this very bad decision to employ a new informed consent mechanism. We never want to make that same mistake again.

Dr. Moertel: I'd like to ask Dr. Simon a question about large study numbers. When we project a study that is going to have exquisitely high power and answer questions within all sorts of subsets, we frequently come up with rather staggering numbers of patients that are going to have to be entered in these trials. These measure literally in the thousands. This, I think, is rather frightening to a lot of us who are trying to approach a study like this. It means an enormous amount of investment over a long period of time, frequently on a very speculative issue. It immediately cuts down the number of groups that are able to approach these areas as well as the number of areas we are able to approach. I wonder if rather than laying out these enormous studies of several thousand patients and perhaps having them never completed, it wouldn't be better to set "extending" rules. These could be comparable to the "stopping" rules we have just heard about. You could start off with a study of relatively modest objectives, relatively low power. Then, based on a modest patient entry and early results, if you feel that there is a reasonable likelihood that this is going to be a successful study that will require a larger patient entry, you can extend it with more enthusiasm. On the other side of the coin, you can abort a lot of studies that really turn out to be bummers and avoid useless expenditure of effort and money.

Dr. Simon: I'd like to make two comments. First, concerning separate analysis of subsets, in many cases it is just not feasible to plan the study to have enough patients in each subset you would like to analyze separately. What I was trying to emphasize was that if you have subsets in which the prognosis is grossly different or if you have strong biological reasons to believe that relative treatment efficacy for patients in different subsets may be totally different, then I think this study needs to be planned to have enough patients to look at subsets separately. Generally it's not practical or strategically sensible to plan a study to look at multiple subsets separately. Second, I think that your comment does raise the very important issue of how one effectively utilizes accumulating information to determine what trials to do and what trials to continue. The extensive statistical literature on sequential analysis has primarily focused on early stopping when you see a large difference. Those kinds of trials are relatively rare. The much more common question is when can you stop early because your results are inconsistent with a clinically substantial difference. I think that's really the same thing as an extension, even though I tend to think of it in terms of having a target sample size sufficient to detect a clinically important difference and stopping early if results are not consistent with a difference of that magnitude. I have published one paper with Gordon Lan and Max Halperin on that topic, and there are others. Susan Ellenberg will be presenting a paper this afternoon that permits one to do just that.

Dr. Moertel: Actually, there are more practical problems involved that I think would push you toward an "extending" rule rather than a "stopping" rule. For example, to get 700, 800, 900 patients on a trial and know that you are going to get them, you really have to start off by setting up some great multigroup endeavor which is tremendously difficult logistically, expensive, and in fact, may not be necessary. Whereas, if you could start off the other way, for example, with SWOG doing a smaller scale study with the idea that given a certain result, you would later extend it to be an intergroup study, it would be vastly easier from a practical standpoint.

Dr. Simon: You have a very good point, though I think there would be some limitations in what you can do in terms of what sort of interim differences are promising enough to make you want to continue. I think your suggestion is interesting and important and warrants careful study.

Analysis and Interpretation of the Comparison of Survival by Treatment Outcome Variables in Cancer Clinical Trials[1]

James R. Anderson,[2,3,*] Kevin C. Cain,[2] Richard D. Gelber,[2] and Rebecca S. Gelman[2,4]

In a previous paper, we examined the common practice of presenting survival comparisons of responders versus nonresponders when reporting the results of the evaluation of chemotherapeutic regimens for cancer patients with advanced measurable disease (1). Here, although we summarize the findings of our previous work and describe recent work by others, we mainly focus our attention on problems involving the interpretation of analyses in which the comparisons of survival are made on the basis of treatment outcome variables (eg, response, toxicity, and percent of protocol dose delivered). The common factor in all of these comparisons is that the category into which a patient falls (responder vs nonresponder, high dose received vs low dose received) is not known at the time the patient enters the study but rather is an outcome of the treatment. The probability of a particular outcome is frequently related to the survival of the patient. Survival time may also be the time period that a patient is at risk of experiencing the outcome used for classification (eg, response, toxicity). In this situation, an analysis which categorizes patients according to the outcome variable and measures survival from study entry will be clearly biased.

Analysis of Survival by Tumor Response Category

In cancer studies involving the evaluation of new drugs or drug combinations for patients with advanced measurable disease, the outcome measures of treatment success most often used are objective tumor response [eg, complete response (CR), partial response (PR), no response, and progressive disease] and survival from the start of treatment. Analysis of survival by tumor response category is often part of the reported results. Patients are usually categorized by best response, and life-table estimates of survival for each response group are calculated, with survival usually measured from the start of treatment. The statistical significance of the observed differences is determined by a log-rank test or another significance test appropriate for censored data. These comparisons are common in the cancer literature; a recent review found 228 articles published in *Cancer* or *Cancer Treatment Reports* presenting data on responders and nonresponders, 140 (61%) of which provided statistical comparisons of these groups and reported *P* values (2). Our own review of the first nine issues of *Cancer* published in 1982 showed that 15 of 31 (48%) reports on the treatment of patients with advanced cancer compared survival for responders and nonresponders, and nine of 15 (60%) reported *P* values for the comparison (1). A more recent review of *Cancer* for 1984 showed that the practice of including such comparisons in clinical reports continues (3–10).

To understand the problems surrounding the analysis and interpretation of comparisons of survival by response category, we first consider the reasons why such analyses are done at all. One justification for performing the analysis is as follows. If it is demonstrated that responders experience longer survival than nonresponders, then it is claimed that response causes longer survival. The fact that response causes longer survival is then used as evidence that more aggressive therapy is warranted, since it will produce a higher response rate, which in turn will result in increased survival.

An alternate argument considers the analysis of survival by response category as a substitute for a placebo-controlled clinical trial. Responders are those who benefited from treatment and thus constitute the treatment group. Nonresponders did not benefit from the treatment, and thus their survival experience can be considered to be equivalent to that of untreated patients (placebo controls). Thus, analysis of survival by response

[1]Supported by Public Health Service grants CA-19589, CA-23415, CA-09337, CA-33601, and CA-23318 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

[2]Department of Biostatistics and Epidemiology, Dana-Farber Cancer Institute, Boston, MA.

[3]Department of Biostatistics, Harvard School of Public Health, Boston, MA.

[4]We thank the Eastern Cooperative Oncology Group for permission to use data from studies EST 2174 and 2177 for illustration and Judy Stephens for her assistance in preparing the drafts of this manuscript.

[*] *Reprint requests to:* James R. Anderson, PhD, Department of Biostatistics, Harvard School of Public Health, 677 Huntingdon Ave, Boston, MA 02115.

category can be considered as a surrogate for the survival comparison of treated and untreated patients.

We will describe why the usual methods used for the comparison of survival by response category are statistically invalid and demonstrate two methods for comparing survival by response that are statistically valid but which do not allow for the testing of the efficacy of treatment.

Analysis of Survival by Response: Statistical Issues

The usual method of comparing responders to nonresponders is wrong and should never be used (1). The survival plots produced are biased, the statistical test of significance is invalid, and conclusions drawn from these analyses are misleading. The bias results from the requirement that responders survive long enough for the response to be observed; no such requirement is necessary for the nonresponders. Note that the usual method attributes the time that responders spend in nonresponse (the period from entering study to response) to the response category. This results in inappropriately favorable survival for the responder group and in inappropriately unfavorable survival for the nonresponder group.

An appropriate statistical test of significance for the null hypothesis that the death rate for responders does not differ from that for nonresponders can be performed by considering response state as a time-varying covariate using the Mantel-Byar procedure (1,11). This allows for a valid comparison of survival by response category, but does not provide estimates of survival probabilities by response category.

Another approach is to select some fixed time after the initiation of therapy as a landmark for conducting the analysis of the effect of response on survival. Patients still alive and under follow-up at the landmark are classified as to whether they had responded before the landmark time. Patients are then followed forward in time from the landmark to ascertain whether survival from the landmark depends on the patient's response status at the landmark. A valid statistical test of significance may be performed, and the landmark method provides meaningful, unbiased conditional estimates of survival (1).

Estimates of survival probabilities computed using ideas from the landmark and Mantel-Byar approach have been proposed by Simon and Makuch (12). The investigators first fix a landmark from which survival will be measured. An estimate of survival for the patients who are nonresponders at the landmark is obtained using the Mantel-Byar model, which considers patients who eventually respond to be censored for survival in the nonresponse category at the time of response. Survival for responders is computed in a similar fashion, allowing nonresponders who eventually respond to enter the response state at the time of their response. It is important to note that the estimated survival curve of responders, unlike that for nonresponders, does not describe the experience of any one group of patients. Instead, it summarizes the probability of surviving beyond Time t, conditional upon a patient being in response at the landmark Time x, or entering the response state between Time x and Time t.

Problems With Interpretation

Even when statistically appropriate analyses of the association between survival and response are conducted, neither of the previously mentioned justifications for the analysis can be supported. It is wrong to imply that longer survival for responders proves that the treatment prolongs survival or that more aggressive therapy is warranted. It is impossible to differentiate these patients from those in whom response acts as a marker for a favorable subgroup of patients. This subgroup may be prognostically favorable because of known or unknown characteristics of the disease (extent, tumor type, tumor markers) and/or of the patient (age, sex, race). Thus, responders may survive longer than nonresponders, not because of an effect of response on survival but because response identifies patients with pretreatment characteristics which favor longer survival.

Example

We use data from sequential Eastern Cooperative Oncology Group (ECOG) studies of premenopausal women with metastatic breast cancer. These analyses are for illustration only. In study EST 2174, premenopausal women with progressive measurable metastatic breast cancer were initially treated with oophorectomy. Patients with at least stable disease 12 weeks following oophorectomy were then randomized to either cyclophosphamide, methotrexate, and 5-FU or to observation. In study EST 2177, which followed EST 2174 for the treatment of metastatic breast cancer in premenopausal women, one group of patients with estrogen-receptor (ER)-positive or ER-unknown disease was randomized to receive treatment with cyclophosphamide, doxorubicin, and 5-FU (CAF) or to oophorectomy plus CAF. (This latter group is excluded from the present discussion.) Table 1 shows the response rates for initial therapy for patients treated initially with oophorectomy in EST 2174 or with CAF only in EST 2177.

If responders are improperly compared to nonresponders in each of these studies, the results show a statistically significant survival advantage for patients responding to treatment (table 2, figs 1 and 2).

Since responders survive longer than nonresponders and since patients treated with CAF in EST 2177 have a response rate (71%) more than three times greater than that for oophorectomy in EST 2174 (21%), it follows (by the usual argument) that the survival for patients receiv-

TABLE 1.—Response rates

EST No.	Treatment	No. of patients	% response				% unevaluable
			CR	PR	No change	Progressive disease	
2174	Oophorectomy	105	5	16	41	33	5
2177	CAF	95	27	43	18	3	8

ing CAF will be better than that for patients receiving oophorectomy.

However, there is no difference in survival experience between EST 2174 and the CAF arm of EST 2177 (table 3, fig 3).

A statistically valid analysis of survival by response category for each study provides some insight into the apparent contradiction. Survival was compared using the landmark method; for each study patients alive at 3 months were classified as to whether a response had occurred before 3 months. Then patients were followed forward from the 3-month landmark. This analysis addresses the question of whether patients alive at 3 months who had previously responded to treatment have a subsequent survival experience which is different from those patients alive at 3 months who had not previously responded. This analysis was repeated using 6 months as the landmark.

EST 2174

All 22 responding patients on EST 2174 did so before the 3-month landmark and all were alive at 3 months. Eight patients who did not respond to treatment died before the 3-month landmark. Thus, 97 patients (22 previous responders, 75 with no previous response) were followed forward from the 3-month landmark and their survival experience was compared. Those patients responding before 3 months had a significantly more favorable survival experience than those not responding before 3 months ($P = 0.005$, fig 4). This difference in survival experience was also observed when 6 months was used as a landmark. All 22 responders were still alive and their survival from 6 months was compared to that of the 70 patients without response who were still alive at 6 months ($P = 0.01$, fig 5). These results demonstrate an association between response and survival, but cannot be used to imply that response to oophorectomy "causes" increased survival. An alternative explanation may be proposed. One might hypothesize that the tumors responding to oophorectomy are strongly ER-positive, and thus estrogen-dependent tumors. Premenopausal women with stage II ER-positive tumors have a more favorable prognosis than those with tumors that are not ER-positive (13). Thus, it is possible that the response to oophorectomy is a marker for patients with strongly ER-positive tumors and that the comparison of responders versus nonresponders simply reflects a more favorable prognosis for these patients generally.

TABLE 2.—Survival comparison of responders to nonresponders

EST No.	Treatment		No. of patients	2-yr survival	
2174	Oophorectomy	Responders	22	82%	
					$P = 0.002$*
		Nonresponders	83	42%	
2177	CAF	Responders	67	59%	
					$P = 0.0009$†
		Nonresponders	28	31%	

*Fig 1.
†Fig 2.

EST 2177: CAF-Only Patients

The landmark analyses of survival by response prior to 3 and 6 months were also conducted for EST 2177. The status of patients treated with CAF only on EST 2177 at the 3-month landmark was as follows: no prior response, 43 (includes 20 subsequent responders); prior response, 47; and died, five (all nonresponders).

A comparison of survival from the 3-month landmark between the 47 patients who had achieved a prior response and the 43 patients who had not achieved a prior response showed no difference in survival ($P = 0.84$, fig 6).

The status of patients at the 6-month landmark was: no prior response, 29 (includes eight subsequent responders); prior response, 58; and died, 14 (one responder, 13 nonresponders).

A comparison of survival from the 6-month landmark between the 58 patients who had achieved a prior response and the 29 patients who had not achieved a prior response showed no difference in survival ($P = 0.67$, fig 7).

TABLE 3.—Survival

EST No.	Treatment	No. of patients	2-yr survival	
2174	Oophorectomy	105	50%	
				$P = 0.97$*
2177	CAF	95	51%	

*Fig 3.

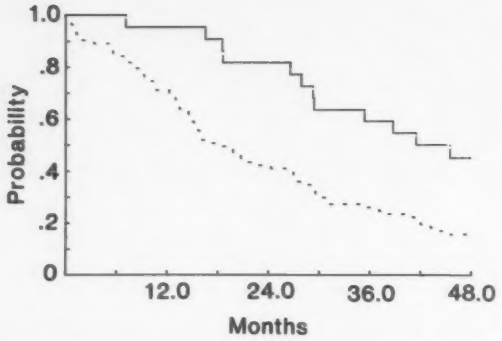


FIGURE 1.—Comparison of survival of responders (solid line) and nonresponders (dashed line) (EST 2174) by the usual method, $P = 0.002$.

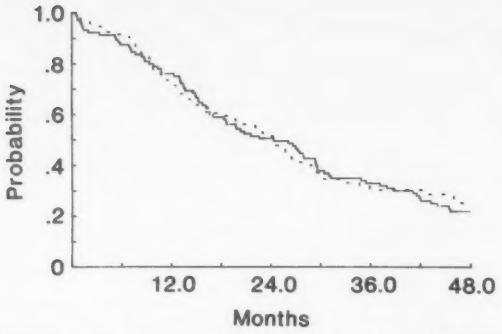


FIGURE 3.—Comparison of survival of patients treated on protocol EST 2174 (solid line) and of patients treated on EST 2177 (dashed line), $P = 0.97$.

These analyses provided no evidence that the survival of patients alive at the landmark, having responded to therapy before these landmarks, was any different from the survival of those patients alive, having not responded before these landmarks. Thus, it appears that the achievement of a response was not associated with survival time for patients in this study.

DISCUSSION

As we and others have demonstrated, the usual method of comparing responders and nonresponders in cancer clinical trials is wrong and should not be used (1,2,12,14-17). These difficulties in analysis and interpretation exist whenever survival comparisons are made using categories based on outcomes of treatment. Skeel et al (18) compared survival of children with acute lymphoblastic leukemia (ALL) and claimed that an association of highest percentage of bone marrow lymphocytes with prolonged survival implied that "the occurrence of longer remissions and survival times in the group with lymphocytosis in ALL may, in fact, indicate an important relationship between this type of cellular response and successful control of the leukemic illness." Breslow and Zandstra (19) showed that this association was a statistical artifact; the larger the number of observations made on any one patient, the greater the chance that any one observation will exceed a specified level simply by chance. Thus, those patients with long remissions had higher maximum lymphocyte counts, or, conversely, those with lymphocytosis were more likely to have longer remissions. Redmond et al (20) have discussed the methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease-free survival. They note that in this context (as well as with other analyses of "time to failure" comparing groups defined by outcomes of treatment) it is difficult, if not impossible, to distinguish cause from effect.

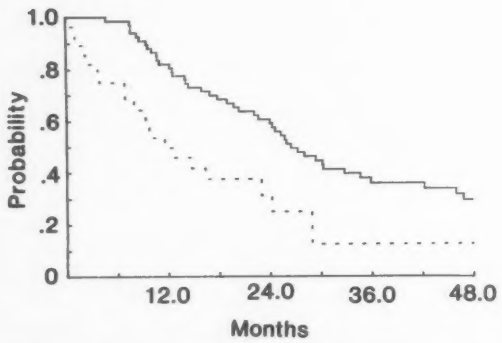


FIGURE 2.—Comparison of survival of responders (solid line) and nonresponders (dashed line) (EST 2177) by the usual method, $P = 0.0009$.

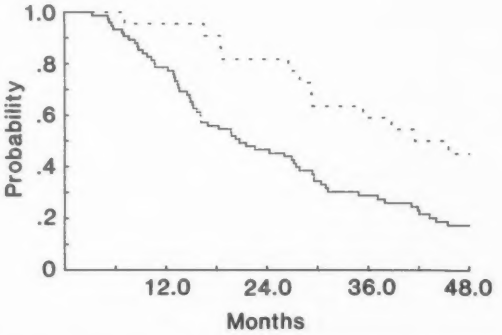


FIGURE 4.—Comparison of survival by response category by landmark method for patients treated on EST 2174; patients with no prior response at 3 mos (solid line), patients with prior response at 3 mos (dashed line), $P = 0.005$.

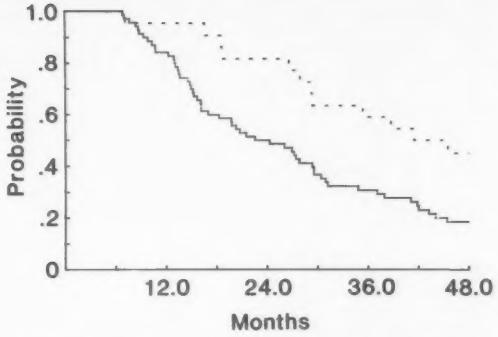


FIGURE 5.—Comparison of survival by response category by landmark method for patients treated on EST 2174; patients with no prior response at 6 mos (solid line), patients with prior response at 6 mos (dashed line), $P = 0.01$.

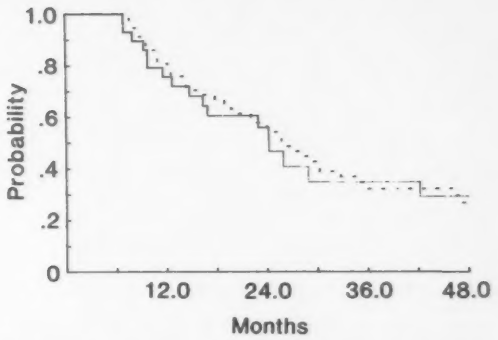


FIGURE 7.—Comparison of survival by response category by landmark method for patients treated on EST 2177; patients with no prior response at 6 mos (solid line), patients with prior response at 6 mos (dashed line), $P = 0.67$.

We have described two statistically appropriate methods for assessing differences in survival by response. However, we emphasize again the difficulties in interpretation of the results of even appropriately conducted analyses. It is wrong to imply that longer survival for responders proves that the treatment has prolonged survival or that more aggressive therapy is warranted. This is true for other comparisons of survival by treatment outcome variables. The fact that a relationship between survival and another outcome variable has been demonstrated in an appropriately conducted analysis does not necessarily imply a causal relationship. An outcome variable such as percent of protocol dose delivered or toxicity experienced may be associated with survival because some basic underlying factor (extent of disease, health of the patient) is strongly associated with both survival and the occurrence of the outcome of interest. Thus, the outcome variable is merely identifying those more prognostically favorable patients who will have better survival in any case.

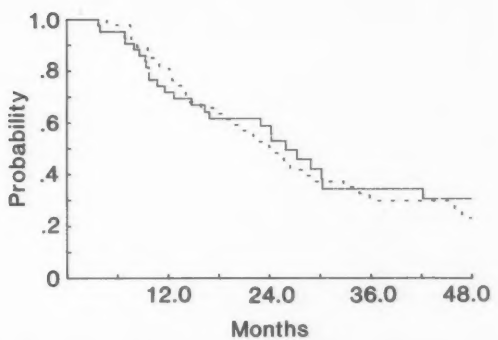


FIGURE 6.—Comparison of survival by response category by landmark method for patients treated on EST 2177; patients with no prior response at 3 mos (solid line), patients with prior response at 3 mos (dashed line), $P = 0.84$.

While plots of survival by tumor response status based on the landmark method may be useful descriptively, we find no justifiable reason for any survival comparison of responders versus nonresponders. Publication of even statistically appropriate comparisons of survival by response category should be strongly discouraged. *Cancer Treatment Reports* will no longer publish papers which include comparisons of survival by tumor response (21). We encourage editors of other journals in which this has been the practice in the past to do the same.

REFERENCES

1. ANDERSON JR, CAIN KC, and GELBER RD. Analysis of survival by tumor response. J Clin Oncol 1:710-719, 1983.
2. WEISS GB, BUNCE H, III, and HOKANSON JA. Comparing survival of responders and nonresponders after treatment: a potential source of confusion in interpreting cancer clinical trials. Controlled Clin Trials 4:43-52, 1983.
3. STEPHENS RL, VAUGHN C, LANE M, ET AL. Adriamycin and cyclophosphamide versus hydroxyurea in advanced prostatic cancer. A randomized Southwest Oncology Group study. Cancer 53:406-410, 1984.
4. BRODOVSKY HS, BAUER M, HORTON J, ET AL. Comparison of melphalan with cyclophosphamide, methotrexate, and 5-fluorouracil in patients with ovarian cancer. Cancer 53:844-852, 1984.
5. SAMSON MK, RIVKIN SE, JONES SE, ET AL. Dose-response and dose-survival advantage for high versus low-dose cisplatin combined with vinblastine and bleomycin in disseminated testicular cancer. A Southwest Oncology Group study. Cancer 53:1029-1035, 1984.
6. KISH JA, WEAVER A, JACOBS J, ET AL. Cisplatin and 5-fluorouracil infusion in patients with recurrent and disseminated epidermoid cancer of the head and neck. Cancer 53:1819-1824, 1984.
7. PINEDO HM, BRAMWELL VHC, MOURIDSEN HT, ET AL. Cyvadic in advanced soft tissue sarcoma: a randomized study comparing two schedules. A study of the EORTC Soft Tissue and Bone Sarcoma Group. Cancer 53:1825-1832, 1984.

8. PINNANANENI K, YAP H-Y, BUZDAR AU, ET AL. Adriamycin, dibromodulcitol, and mitomycin combination chemotherapy for patients with metastatic breast carcinoma previously treated with cyclophosphamide, methotrexate, 5-fluorouracil, vincristine, and prednisone. Cancer 53:1841-1844, 1984.
9. AL-KATIB A, KOZINER B, KURLAND E, ET AL. Treatment of diffuse poorly differentiated lymphocytic lymphoma. An analysis of prognostic variables. Cancer 53:2404-2412, 1984.
10. KOZINER B, SKLAROFF R, LITTLE C, ET AL. NHL-3 protocol. Six-drug combination chemotherapy for non-Hodgkin's lymphoma. Cancer 53:2592-2600, 1984.
11. MANTEL N, and BYER DP. Evaluation of response-time data involving transient states: an illustration using heart transplant data. JASA 69:81-86, 1974.
12. SIMON R, and MAKUCH RW. A non-parametric graphical representation of the relationship between survival and the occurrence of an event: application to responder versus non-responder bias. Stat Med 3:35-44, 1984.
13. FISHER B, REDMOND C, BROWN A, ET AL. Influence of tumor estrogen and progesterone receptor levels on the response to tamoxifen and chemotherapy in primary breast cancer. J Clin Oncol 1:227-241, 1983.
14. MANTEL N. An uncontrolled clinical trial—treatment response or spontaneous improvement? Controlled Clin Trials 3:369-370, 1982.
15. MANTEL N. Responder versus nonresponder comparisons: daunorubicin plus prednisone in treatment of acute nonlymphocytic leukemia. Cancer Treat Rep 67:315-316, 1983.
16. TANNOCK I, and MURPHY K. Letter. J Clin Oncol 1:448, 1983.
17. OYE RK, and SHAPIRO MF. Reporting results from chemotherapy trials. Does response make a difference in patient survival? JAMA 252:2722-2725, 1984.
18. SKEEL RT, HENDERSON ES, and BENNETT JM. The significance of bone marrow lymphocytosis of acute leukemia patients in remission. Blood 32:767-773, 1968.
19. BRESLOW N, and ZANDSTRA R. A note on the relationship between bone marrow lymphocytosis and remission duration in acute leukemia. Blood 36:246-249, 1970.
20. REDMOND C, FISHER B, and WIEAND HS. The methodologic dilemma in retrospectively correlating the amount of chemotherapy received in adjuvant therapy protocols with disease-free survival. Cancer Treat Rep 67:519-526, 1983.
21. SIMON R, and WITTES RE. Methodologic guidelines for reports of clinical trials. Cancer Treat Rep 69:1-3, 1985.

DISCUSSION

Q: (Dr. Tom Reichert, Becton-Dickinson) Jim, I find this a very fascinating presentation. But I'm a little troubled by the methodology that you put forward. The usual comparison has clearly got its problems—but the two methods you've proposed do not really address the question of interest. The landmark method compares two cadres, one of which is the group of patients who have responded by a landmark time; the other group is a cadre that has failed to respond. That's quite different from the original objective: comparing the people who respond to therapy and the people who do not respond to therapy. Similarly, the Mantel-Byar test provides us with a single survival curve of patients who are censored both by response and by death—also very different from the item of interest. I'm reminded of the anecdote of the drunk looking for his keys under the streetlight; what we want is over there in the dark, and I'm afraid it's quite lost, and these procedures that we know how to do are over here under the light. They really don't help us to do precisely what it is that we set out to do and I think that it's misleading to suggest that they provide an answer to the question at hand, which is that we'd like to know how to compare responders and nonresponders.

A: (Dr. Anderson) Two comments. First, in an attempt to provide statistically valid methods of comparing patients' survival by response, one is forced into the situation of making certain restrictions to assure statistically valid tests. The second comment I would make is that it may be that that which is in the dark should remain in the dark. It seems to me that none of the comparisons of survival by responders and nonresponders that I've seen proposed can be unconfounded with situations where differences can be explained by other mechanisms. And so, even if a method of comparing them to suit you could be found, there is no way to differentiate the situation that people are interested in, where response has made a difference, from the situation where response is simply a marker for prognostically favorable patients.

Q: (Dr. Reichert) I think the important thing, though, is to avoid investing us with new dogma about procedures when they may not, in fact, address the question the answer to which we've all been searching.

A: Well, the hypotheses are clearly stated. If those are the questions of interest, namely, if one is interested in assessing the relative risk of death among those with prior response versus those without prior response, then a Mantel-Byar procedure addresses that question. I agree with you that that may be different from the question of survival by response category, but it's a properly formulated hypothesis that can be tested appropriately. Likewise, the landmark method allows one to investigate a particular type of question. It may not be the question of interest, but it's clear that the usual method doesn't address the usual question of interest, and that the results from that analysis are wrong and misleading.

Q: (Dr. Joe Pater, Kingston) It seems to me you've defined a solution to a problem that we're really not trying to face. Investigators, especially physicians who are trying to prescribe drugs, are interested in the question, Does the administration of this treatment prolong life, as compared to not administering this treatment? And this solution doesn't really address that question. Nobody really knows whether the use of combination chemotherapy prolongs survival in breast cancer, but if there is any evidence that it does, it's in those patients who are about to die very quickly and who are saved from dying very quickly by the administration of treatment. Those people would be neglected by your method and you would have a false-negative result by using the landmark method or the Mantel-Byar method applied to that question.

A: I would argue that if the question of interest is whether single-agent therapy or combination chemotherapy benefits patients in terms of long-term survival, the only way to address that question appropriately is to do a study of that—which means one group of patients gets the therapy and another group of patients is randomly allocated to get no therapy or something different. Then one has statistically appropriate methods for assessing whether differences exist. It is inappropriate, for the reasons that I mentioned, to use analysis by response as a surrogate for that.

(Dr. Pater) I couldn't agree more. It's just a question of whether another method might be seen by some naive people as a solution to the problem which could substitute for a randomized trial, which I agree is the only solution to the problem.

A: With that, I'm in perfect agreement.

Q: (Dr. Ron Christensen, Entropy Limited) I wish to raise a statistical question regarding the landmark method that Dr. Anderson mentioned. It's easy to show that it is biased in exactly the opposite direction than the usual method is biased. The landmark method drops early deaths; therefore, in any expected value theory, which is what you're working with in calculating bias, it is going to be dropping a larger than expected fraction of nonresponders. Therefore, in any normal clinical application, it would be biased in favor of nonresponders, just in the way that the usual procedure is biased in favor of responders. So I don't see that it's surprising at all that you'll push the curve in the opposite direction by using the landmark method. I think the landmark method should be just as cautiously used as the usual method in any applications, because of its own bias.

A: I did try to make the point that the application of the landmark method involves conditional analysis, both on those individuals alive at the landmark and those individuals' response state at the landmark. Given that that is where one begins, the analysis using the landmark method is not biased. I agree that if your frame of reference is back at time zero, the landmark method doesn't address that question.

Q: (Dr. Charles Moertel, Mayo) I've heard these same arguments many times before, regarding the inadequacies of drawing inferences because responders live longer than nonresponders, and I doubt if there are very few in this room, at least, that would care to carry the banner for such analyses as indicating therapeutic activity that adds to life span. And yet, as we have heard, and as we repeatedly see, such analyses are constantly being presented in our journals, with the clear purpose of creating an impression that a therapy is effective in prolonging life span because responders live longer than nonresponders. I'm certain that we can pick up any oncology journal, this month's issue, and find those same sorts of curves displayed. Dr. Wittes, how would you like to take the leadership in this area by refusing to accept those curves for publication in *Cancer Treatment Reports*?

A: (Dr. Robert Wittes, NCI) Thank you, Dr. Moertel, I'm glad you asked. We do refuse to accept such curves in *Cancer Treatment Reports*.

An Efficient Design for Phase III Studies of Combination Chemotherapies

Susan S. Ellenberg* and Mario A. Eisenberger[1]

Demonstration of antitumor activity is an essential early element of the investigation of potential anticancer agents. Activity is defined by tumor shrinkage in a minimally acceptable proportion of patients. Studies in which the goal is to establish antitumor activity are referred to as phase II clinical trials.

When multiple active drugs have been identified within a particular disease setting, the investigation of combinations of these drugs proceeds on the assumption that simultaneous use of multiple agents will improve the probability of achieving tumor regression. Initial small uncontrolled studies of combination regimens can be justified on the basis of ensuring an acceptable level of toxicity. Demonstration of activity is generally not a concern since the individual drugs used will have already been shown to be active. Although estimation of the actual level of activity is often given as a justification for these uncontrolled studies, the size of such studies is usually too small (20–30 patients) to provide a reasonably precise estimate of the true response rate. For example, the 90% confidence interval around an observed response rate of 30% in a study of 30 patients extends from 15% to 65%. With this type of result it is difficult for an investigator to decide whether or not the new combination is worthy of further study. In some cases, the results are so clearly negative or positive that the decision (at least for that investigator) as to whether to continue study of the regimen is relatively easy. In many cases, however, the results are in that broad middle range in which interpretation is difficult.

Despite the imprecision of small uncontrolled phase II studies, many such studies of combination regimens are carried out. Commonly stated goals of such phase II studies are to "confirm" or "further investigate" the previously reported results. These individual studies tend to demonstrate a wide range of outcomes, consistent with the wide confidence intervals which could be constructed about any single result. An example is the series of phase II studies of FAM (5-FU, doxorubicin, and mitomycin) in advanced gastric cancer, in which response rates ranging from 18% to 65% have been reported (1–9).

One reason for the reluctance of many investigators to initiate controlled studies of combination regimens is certainly the large sample size required. Comparative controlled trials, or phase III trials, are designed to determine, with low probabilities of error, whether a new treatment differs in effect from a control treatment. Such studies may require ten or more times the number of patients conventionally entered in a phase II study. An investigator may not want to commit such resources until uncontrolled studies indicate that this new regimen is likely to be a substantial improvement over single-agent or other "standard" therapy. Ironically, this strategy can make it difficult to get a potentially definitive randomized study off the ground, if the initial uncontrolled trial or trials appear highly positive. However, the major problem with this strategy is that promisingly high response rates in small uncontrolled studies are often found to be overestimates when phase III studies are performed. Much speculation has been generated regarding the possible causes of this phenomenon: selection of patients, differing response criteria, "regression to the mean," among others. A good example can be found in the recent literature of head and neck cancer. A summary of data from chemotherapy trials in advanced head and neck cancer is shown in table I. These data originated with a literature review of controlled studies of combination therapy for recurrent and/or metastatic head and neck cancer in which single-agent methotrexate was the control arm (10). In the first column, data from all uncontrolled studies of experimental regimens referred to in these articles as the "stimulus" for doing the randomized controlled trials are summarized. The second column gives data on these same regimens obtained from the randomized trials. The third column shows data on the control regimen, methotrexate, in the randomized trials. The average response rate for experimental regimens in the uncontrolled studies was 45%, 16 points higher than the aver-

[1]Cancer Therapy Evaluation Program (*S. S. Ellenberg*) and Clinical Investigation Branch (*M. A. Eisenberger*), Division of Cancer Treatment, National Cancer Institute, Bethesda, MD.

**Reprint requests to:* Dr. Susan S. Ellenberg, Cancer Therapy Evaluation Program, Division of Cancer Treatment, National Cancer Institute, Landow Bldg, Rm 4B06, National Institutes of Health, Bethesda, MD 20205.

TABLE I

RESPONSE RATES FOR EXPERIMENTAL REGIMENS IN PHASE II AND PHASE III STUDIES

Experimental Regimen	PHASE II Number of Studies	PHASE II Overall Response Rate	PHASE III Number of Studies	PHASE III Overall Response Rate
High-dose Methotrexate + Leucovorin Rescue	9	93/168 (55%)	5	53/197 (27%)
Cisplatin	6	50/186 (27%)	2	10/71 (14%)
Cisplatin, Vinblastine Bleomycin	2	29/52 (56%)	1	22/92 (24%)
Cisplatin, Vincristine Bleomycin	1	17/27 (48%)	1	11/27 (41%)
Cisplatin, Methotrexate, Bleomycin	1	19/31 (61%)	1	39/81 (48%)
All Regimens	19	208/464 (45%)	10	135/468 (29%)

age response rate for these same regimens in a controlled setting. The average response rate for methotrexate in the controlled trials was 27%, compared to 29% for the experimental regimens; in one-half of the randomized studies, methotrexate produced the higher response rate.

Although most experimental investigators are aware of the unreliability of data obtained in small uncontrolled studies, they are concerned about committing themselves to lengthy studies of new regimens which appear early on to lack benefit. A commitment to accrue the ≥ 150 patients required by a comparative study may look like an unfortunate decision if early results show no indication at all that the experimental therapy offers any improvement over standard therapy. An investigator in this situation might consider abandoning the study, even though only a small number of patients have been evaluated.

We have investigated the properties of a study design which would incorporate the intuitively appealing concept of early termination in the face of initial negative results. This investigation has shown that such early stopping rules can indeed be employed in phase III studies without substantially affecting the reliability of the conclusions to be drawn from the trial's results. The particular early stopping rule to be described is arbitrarily chosen on the basis of simplicity of application.

Consider the following two-stage plan:

Stage 1.—Observe n patients on each arm. If the number of responses to experimental therapy B is the same as or fewer than the number of responses to standard therapy A, then terminate the trial and discard B as a therapy of interest. Otherwise, proceed to the second stage.

Stage 2.—Enter the remaining patients required according to fixed sample size considerations and perform the usual final analyses.

The obvious advantage of such a design is that fewer patients may be required. The obvious concern is that studies which would eventually show benefit for a new regimen may be stopped incorrectly. It will be shown that early stopping rules can be constructed so that the problem of incorrectly stopping potentially positive studies is minimized while still providing substantially reduced expected sample sizes when insufficiently effective regimens are studied.

To implement this proposed two-stage design we need to determine the value of n, the minimum number of patients to be studied before making a judgment about stopping or continuing. A reasonable way of determining n is to find that point in the trial at which the observation of equivalent or inferior outcomes in the experimental group as compared with the control group (the early stopping criterion) is inconsistent with the prespecified improvement of interest. In other words, what n can we choose to make the probability of erroneously stopping early acceptably small? We find n by first defining γ as this "acceptably small" probability of erroneously stopping early, and then finding the smallest value of n for which the following inequality holds:

$$\sum_{j=0}^{n} \binom{n}{j} P_A^{\,j}(1-P_A)^{n-j} \sum_{k=0}^{j} \binom{n}{k} P_B^{\,k}(1-P_B)^{n-k} \leq \gamma,$$

where P_A and P_B are the (presumed) true response rates to standard therapy A and experimental therapy B, respectively. Examples of values of n for a variety of values of P_A and P_B, and for γ arbitrarily set at 0.05, are given in table II. The fixed sample size requirements for

two different levels of power are also given, to provide some idea of the degree to which we could reduce the number of patients entered in a trial of an experimental therapy.

A desirable property of an early stopping rule would be that the probability of stopping early would be high when the true treatment benefit was small or zero (or worse) and would be low when the true treatment benefit was great. In figures 1-3, for detection of small, moderate, and large differences, we see that our proposed design does have this property. The values on the horizontal axis are the *true* values of the experimental response rate. In each case we see that if the new therapy really is equivalent to control, the probability of early stopping is a little more than 50%. If it is worse, there is a high probability of early stopping; if it is better, there is a lower probability of early stopping. If the true rate is in fact the prespecified rate of interest, the probability of early stopping is exactly γ (in this case 5%).

Implementation of this design could result in substantially fewer patients required for study. Table III shows the expected sample sizes per arm, under the assumptions that the standard treatment produces a 20% response rate and for a range of response rates for the experimental therapy. Results for small, moderate, and large differences are given. It should be noted that this design will never result in any of these "expected sample sizes" in a particular trial. Only the fixed sample size or the early stopping point can be observed. The numbers in this table represent averages over many hypothesized trials. They show that when the experimental regimen being tested is approximately equal in efficacy to the standard, or control therapy, the reduction in patient numbers can be substantial. As the benefit of the experimental regimen increases, the probability of stopping early is (appropriately) reduced, and hence the expected sample sizes become larger, approaching the fixed sample size. These patterns of patient numbers are consistent over the range of conventionally acceptable levels of power.

The positive aspect of this design is a potential reduction in the required number of patients. The cost, as stated previously, is an increase in the probability that we will fail to recognize a true therapeutic gain, or, equivalently, a reduction in power. By using this design, we reduce the probability of rejecting the null hypothesis that experimental therapy is no better than control. But simulations show that these power reductions are negligible, as can be seen in table IV. For the usual levels of power considered acceptable and appropriate, a reduction of no more than one to three percentage points is incurred by incorporating this stopping rule. The magnitude of the power reduction does not depend on the size of the difference to be detected. Of course, since we have selected γ as the proportion of times we will stop (incor-

TABLE II

POTENTIAL EARLY STOPPING POINTS ACCORDING TO DIFFERENCE TO BE DETECTED
(Fixed Sample Sizes* In Parenthesis)

	Control Response Rate							
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8
0.2	77 (176/236)							
0.3	25 (68/76)	110 (250/339)						
0.4	13 (31/40)	32 (74/98)	132 (300/408)					
0.5	9 (20/25)	16 (37/48)	37 (83/111)	143 (325/442)				
0.6	6 (14/17)	10 (22/29)	18 (39/52)	38 (86/115)	143 (325/442)			
0.7	4 (10/13)	7 (15/19)	10 (23/30)	18 (39/52)	37 (83/111)	132 (300/408)		
0.8	3 (8/9)	5 (11/13)	7 (15/19)	10 (22/29)	16 (37/48)	32 (74/98)	110 (250/339)	
0.9	3 (6/7)	3 (8/9)	4 (10/13)	6 (14/17)	9 (20/25)	13 (31/40)	25 (58/76)	77 (176/236)

*80% power/90% power
From Fleiss, J.L. "Statistical Methods for Rates and Proportions" (second ed.) John Wiley & Sons, 1981.

TABLE III

POTENTIAL REDUCTION IN PATIENT NUMBERS BY USE OF STOPPING RULE

A. Control response rate: 20%
Minimum experimental response rate of interest: 30%
Stopping point: 110 patients per arm

	Expected Sample Size Per Arm					
	True experimental response rate					Requirement for fixed sample size trial
Desired Power	.05	.10	.15	.20	.25	
1. 80%	110	112	130	176	221	250
2. 85%	110	113	136	193	251	288
3. 90%	110	113	143	217	291	339

B. Control response rate: 20%
Minimum experimental response rate of interest: 40%
Stopping point: 32 patients per arm

	Expected Sample Size Per Arm						
	True experimental response rate						Requirement for fixed sample size trial
Desired Power	.05	.10	.15	.20	.25	.30	
1. 80%	33	36	42	51	59	65	74
2. 85%	33	37	45	55	65	73	84
3. 90%	33	38	48	61	74	84	98

C. Control response rate: 20%
Minimum experimental response rate of interest: 50%
Stopping point: 16 patients per arm

	Expected Sample Size Per Arm								
	True experimental response rate								Requirement for fixed sample size trial
Desired Power	.05	.10	.15	.20	.25	.30	.35	.40	
1. 80%	17	19	22	25	28	30	33	34	37
2. 85%	17	20	23	27	30	33	36	38	41
3. 90%	18	31	25	29	34	38	41	44	48

rectly) when the true benefit is at least as great as specified, the maximum obtainable power is $100 \times (1-\gamma)\%$; in our examples it is 95%.

The particular design we have been discussing will not be applicable to all randomized trials. It may be used in trials in which the alternatives are one-sided, since the stopping rule is based on differences in only one direction. The outcomes must be observed quickly; the methodology would not generally be appropriate for survival studies. Also, for trials in which the goal is to demonstrate equivalence rather than a benefit of a new therapy, sufficient patients should be studied during the first stage to allow for adequate consideration of the small differences which are important to detect in such studies.

These qualifications not withstanding, we feel that this design would be appropriate for many trials of new cancer therapies.

There are many possible variations of this proposed design. One might construct a stopping rule which would provide for early stopping whenever the experimental therapy was not some specified amount better than standard. We could vary the error probability to allow for more or less stringent criteria for early termination. We might consider incorporating an imbalanced randomization into the design if we had had a substantial amount of

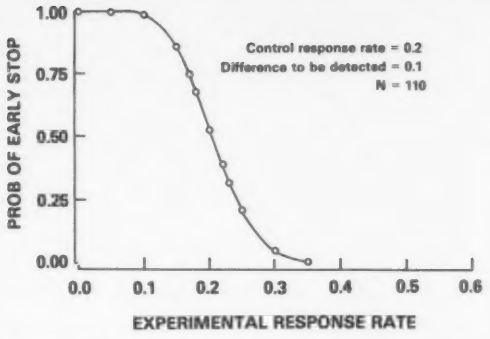


FIGURE 1.—Probability of stopping early as a function of the true experimental response rate. Control response rate = 0.2, difference to be detected = 0.1, potential early stop after 110 patients/arm.

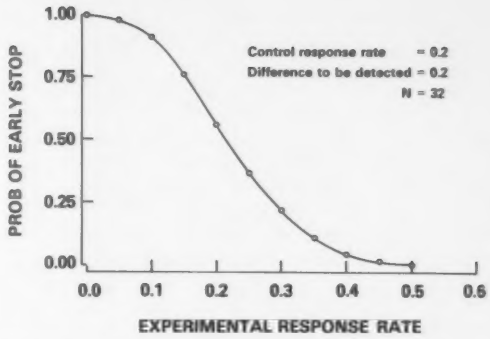


FIGURE 2.—Probability of stopping early as a function of the true experimental response rate. Control response rate = 0.2, difference to be detected = 0.2, potential early stop after 32 patients/arm.

experience with the standard therapy and were anxious to improve our estimates of efficacy and toxicity for the new therapy. An additional decision point (or points) could be built in so that there would be more than one opportunity to terminate early.

In summary, the efficient allocation of patient resources for clinical trials is crucial to progress in development of useful new therapeutic regimens. It follows that investigators will naturally be tempted to abandon a controlled trial, with its large accrual requirement, if negative results are seen early on. The strategy of early stopping in this situation can be justified if the possibility for early stopping is formally built into the study design, so that the error probabilities are controlled. Use of such a design can substantially reduce the numbers of patients required for phase III trials with negligible effect on power.

The concept of a study design which would allow

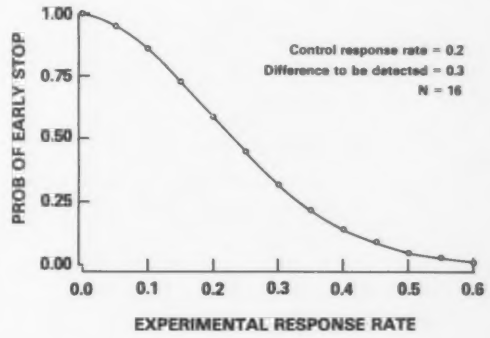


FIGURE 3.—Probability of stopping early as a function of the true experimental response rate. Control response rate = 0.2, difference to be detected = 0.3, potential early stop after 16 patients/arm.

TABLE IV

SIMULATED POWER OF EARLY STOPPING DESIGN

	Simulated Power	
	Desired Power	
Difference of Interest	80%	90%
20% vs. 30%	80%	87%
20% vs. 40%	79%	87%
20% vs. 50%	80%	88%

Based on 10,000 simulations performed via SAS

early termination with negative conclusions when the interim results were not at all promising is not a new one (11-14). However, such designs have not been formally incorporated into study designs of randomized trials of cancer therapy. Stopping rules for early positive results, on the other hand, have received much more attention and are beginning to find their way into clinical trial design. This is understandable, since their development was motivated more by ethical issues than the issue of efficiency to which the stopping rule for negative results is addressed. However, studies which appear negative from beginning to end are far more common than studies which show extremely high levels of benefit. Study designs incorporating rules for early termination on the basis of likely negative results can improve the efficiency of clinical trials and should be seriously considered for future trials of cancer therapy.

REFERENCES

1. MACDONALD JS, SCHEIN PS, WOOLLEY PV, ET AL. 5-Fluorouracil, doxorubicin and mitomycin-C (FAM) combination chemotherapy for advanced gastric cancer. Ann Intern Med 93:53-536, 1980.
2. BITRAN JD, DESSER RK, KOZLOFF MF, ET AL. Treatment of metastatic pancreatic and gastric adenocarcinomas with 5-fluorouracil, adriamycin, and mitomycin-C (FAM). Cancer Treat Rep 63:2049-2051, 1979.
3. HAIM N, COHEN Y, HONIGMAN J, ET AL. Treatment of advanced gastric carcinoma with 5-fluorouracil, adriamycin, and mitomycin-C (FAM). Cancer Chemother Pharmacol 8:277-280, 1982.
4. HAIM N, EPELBAUM R, COHEN Y, ET AL. Further studies on the treatment of advanced gastric cancer by 5-fluorouracil, adriamycin (doxorubicin), and mitomycin-C (modified FAM). Cancer 54:1999-2002, 1984.
5. GILL PG, JONES AM, and ABBOTT R. Chemotherapy of advanced gastric cancer. Aust NZ J Surg 53:239-240, 1983.
6. CUNNINGHAM D, SOUKOP M, MCARDLE CS, ET AL. Advanced gastric cancer: experience in Scotland using 5-fluorouracil, adriamycin and mitomycin-C. Br J Surg 71:673-676, 1984.
7. FORNASIERO A, CARTEI G, DANIELS O, ET AL. FAM2 regimen in disseminated gastric cancer. Tumori 70:77-80, 1984.
8. RIDOLFI R, CASADEI GIENCHI D, CORTESI C, ET AL. A retrospective study of FAM regimen in 38 patients with advanced gastric cancer. Tumori 70:375-379, 1984.
9. FRASCHINI P, BERETTA G, ARNOLD E, ET AL. Confirmed activity of FAM polychemotherapy in advanced gastric carcinomas. Tumori 69:59-64, 1983.
10. EISENBERGER M, POSADA J, SOPER W, ET AL. The current status of chemotherapy. *In* Head and Neck Cancer (Wittes RE, ed). John Wiley & Sons, Ltd, 1985.
11. SPURRIER JD. Two stage sign tests that allow first stage acceptance and rejection. Commun Statist A 7:399-408, 1978.
12. DEMETS DL, and WARE JH. Asymmetric group sequential boundaries for monitoring clinical trials. Biometrika 69:661-663, 1982.
13. GOULD AL, and PECORE VJ. Group sequential methods for clinical trials allowing early acceptance of H_0 and incorporating costs. Biometrika 69:75-80, 1982.
14. LAN KKG, SIMON R, and HALPERIN M. Stochastically curtailed tests in long-term clinical trials. Commun Statist C 1:207-219, 1982.

DISCUSSION

Q: (Dr. Robert Makuch, NCI) Susan, I was wondering if your work fits in with some of the work that Max Halperin did several years ago with stochastically curtailed testing. At some particular fixed point in time during a study you want to calculate the expected probability of getting a result in one particular direction, given the results so far. Does that connect with your work in any way?

A: (Dr. Ellenberg) Well, certainly the philosophy is the same. You want to stop early if it's unlikely that you're going to come out with a positive result in the end, and I think that's the basis of this procedure. Although it's not formally based on a conditional power argument, I think the overall effect is the same. But the mathematical derivation is different.

Q: (Dr. Tom Fleming, Seattle) Philosophically, this approach is related to some of the group sequential designs that we and some others have looked at. It's very intuitive and very appealing from that respect. Some of the other methods that exist do allow for this two-stage or early termination, even if you had survival as a major endpoint. I suggest that actually what you're doing could easily be generalized as well to be used when survival is the endpoint. If you were looking at two survival curves and you were looking at the hazard rate, for example, your rule might say, stop early if the survival curves are equal or worse, that is, if you had a hazard rate of 1 or less. So those ideas could be generalized to the survival situation. You don't really have to wait for a lag time for the rest of the people that have been accrued to see what's going to happen—you can just analyze right at that point. I think these methods would work, with some minor modifications, there as well.

Q: (Dr. Mitchell Gail, NCI) I'd like to second that comment, but with one caveat, and that is that what you're doing is drawing an inference based on the early survival events, and it is quite possible that the long-term survival experience may not reflect the early survival experience. You're relying upon something like a proportional hazards assumption when you consider these methods for survival studies.

A: I agree with both comments. While one could certainly develop these types of stopping rules for testing differences in hazard rates, I agree with Dr. Gail that one has to be more cautious in extrapolating from early survival results than from early response results.

Q: Susan, you look like you got something out of nothing—no loss of power—and I suspect that somewhere around half the design difference is a substantial loss of power. Have you looked at that? I was also wondering if you were going to investigate those things on your last slide—changing gamma and so on.

A: Well, if the true difference were half the design difference, of course you'd have much smaller power even if you carried it out to the planned sample size. I'm not sure that you would get any reduction in power; in other words, if you were looking for a 20 percentage point difference and you had 75 patients per arm, but really the true difference was only 10 percentage points or 20% versus 30%, then your power to detect that difference is not going to be 80%, it's going to be less than 50%. And my guess is that it would not be substantially reduced by stopping early, but I have not looked at that.

Q: (Dr. Satya Dubey, FDA) Statistical quality control methods deal with similar sampling designs. Double sampling and multiple sampling plans generally do not have as good power as a sequential design. Sequential sampling is considered to be most efficient in terms of expected number of patients you will need before you will make correct decisions with a desired power. Consequently, I'm wondering to what extent you had a chance to compare these widely available designs with what you are proposing here.

A: As I said, I have no doubt that there are other procedures which may, in fact, be more

efficient than this one, and one thing that we're going to be interested in doing is to compare a number of these different procedures. The one advantage of a procedure like this, even though it may not be as efficient as something else, is that it's very simple, it's very easy to apply, and if it turns out that it's not so much less efficient than some of the others, it may have a certain appeal.

Q: (Dr. Richard Simon, NCI) There is considerable evidence that two- or three-stage sampling plans do almost as well as fully sequential plans.

Q: (Dr. Daniel Kisner, San Antonio) I think your comment that you may be saving patients would be true only if we could count on a significant number of new combination therapies being better. I think that with the assumption that most new drug combinations will be no better than what we have as standard therapy, the high likelihood is that we are doubling the number of patients it will take to prove that.

A: That's a very interesting and important point. In part, the savings are going to depend on the proportion of new regimens that come out that are useful. If 95% of the new combinations which are tried looked terrible even in the phase II uncontrolled studies and only 5% of the new combinations go on to phase III, then I agree that you would be losing patients by insisting that we don't do any phase II studies of combinations and start out all in controlled trials. On the other hand, if it's the other way around, if most of the combination regimens do, in fact, look good enough in phase II studies to be carried on to phase III, then you do save. So I think that's something that needs to be considered. In any case, we may not need as many phase II studies of some of these combinations prior to going into phase III studies as we have, and perhaps using some kind of design like this may encourage people to start the randomized trial earlier, instead of doing one more phase II study.

Q: Just one comment on the practical implications of doing this. We find, in many cooperative groups, a resistance to stopping studies early, unless there are ethical reasons to stop, such as toxicity or large treatment differences. There's a lot of resistance to stopping when there are small treatment differences midway through the trial. One reason that is given is that we don't have any replacement arms to substitute at that point. If we stop the study at that time, we will not get case credits for our patients anymore. This is a real practical problem, and I think it's going to be faced by us for a long time, unless NCI makes some kind of policy decisions.

A: I don't really think that's a problem, because if you were going to implement a procedure like this, you'd set it up so that you have replacement arms ready, given that you knew you might stop early. Right now, many groups are setting up master protocols for phase II studies where they know as soon as they accrue 30 patients the other arm comes right in. You would have the same kind of set-up for something like this.

(Dr. Robert Wittes, NCI) Could I just make one or two comments about that? It may be possible for us to devise a system by which you get Group credit for stopping studies that deserve to be stopped. The second comment relates to the expected very large increase in the number of new drugs and biologics that are going to be brought into clinical trials over the next few years. There's going to be no shortage of things to test, and I think one of the real merits of exploring the whole early stopping area is that it will enable answers to be gotten with the smallest possible expenditure of patient resources.

A: I would also be amazed to find investigators resistant to stopping a study with early results of the kind that I've showed. In my experience, people are very anxious to be given some kind of rationale for stopping a study like that.

Q: (Dr. Andrew Coldman, Vancouver) I was faced with a somewhat similar problem to this in a trial that stopped rather early because of funding problems and people were looking at the results. And what I did in that case was to put a posterior distribution on the difference. I wonder if you have any comments on a Bayesian approach to giving information back to investigators.

A: I think that would be another interesting approach to consider. As I said, I haven't started to compare the properties of any of these approaches, but I certainly think that could be a reasonable way to approach the problem as well.

Food and Drug Administration Requirements for Approval of New Anticancer Drugs

John R. Johnson* and Robert Temple[1]

For approval of new drugs, the law requires a minimum of two independent well-controlled studies for each indication. The preferred study design is prospective, randomized, and comparative. The use of historical controls, while sometimes acceptable, must be justified. A favorable effect on survival and/or quality of life is generally required for approval. The study of cancer drugs in combination presents problems which are discussed. Drug firms are strongly urged to discuss the protocols with the Food and Drug Administration prior to initiating phase III comparative studies to be used for approval of the drug. [Cancer Treat Rep 69:1155–1157, 1985]

Approval of any new drug requires, under the Food, Drug, and Cosmetic Act and Food and Drug Administration (FDA) regulations, a minimum of two independent well-controlled clinical studies demonstrating the drug's safety and efficacy for each proposed indication. This requirement is in accord with the accepted scientific principle of replicability.

In rare instances new drugs have been approved on the basis of one multicenter study, usually because the results of the study were so striking that repetition would raise serious ethical problems. To be convincing, such a multicenter study should be of prospective randomized design, and results across centers should show appropriate consistency. The results also must be clear and convincing and must show an important therapeutic gain. For cancer drugs, this means cure or probable cure in some of the study patients.

STUDY DESIGN

The required two independent well-controlled studies should, if at all possible, be of randomized prospective design. Historical controls are, for well-recognized reasons, definitely inferior to concurrent controls, and use of historical controls must be justified (1-3). Ethical considerations, overwhelmingly favorable results, or low incidence of the disease may be acceptable justifications.

In cancer drug trials, positive control designs are often used where the test drug is compared with another cancer drug rather than a placebo. Positive control designs are often unavoidable when some treatment appears ethically required, but these studies pose major difficulties when the test drug is not superior to the control drug, because the effectiveness of the control drug is often not well documented in the population to be studied.

PARAMETERS OF EFFICACY FOR ANTICANCER DRUGS

Studies of Advanced Disease

Parameters of efficacy for advanced metastatic disease are survival, quality of life (performance status or pain), objective tumor response, duration of objective tumor response, and time to treatment failure. Survival and quality of life are the key efficacy parameters. In the past, new anticancer drugs were approved solely on the basis of objective tumor response, but this is no longer the case. Biological activity (tumor shrinkage) and efficacy (patient benefit) are not necessarily synonymous. The requirement for survival and quality of life data was initiated on the recommendation of the FDA Oncology Drugs Advisory Committee. Some new drug applicants have complained that the requirement for an effect on survival is inconsistent with FDA requirements for drugs intended for nonmalignant diseases, such as lipid-lowering drugs. These are matters in which disagreements are understandable, but in general we seek survival data where it is the true end point and is reasonably obtainable. Certainly, survival

[1]Office of Drug Research and Review, Center for Drugs and Biologics, Food and Drug Administration, Rockville, MD.

Reprint requests to: John R. Johnson, MD, Rm 9B-45, HFN-150, Parklawn Bldg, 5600 Fishers Lane, Rockville, MD 20857.

is the parameter of primary interest to cancer patients and, unfortunately, it does not take long to accrue survival data in most types of advanced metastatic cancer. It seems difficult to justify use of so toxic a class of drugs if studies fail to show a favorable effect on survival or patient comfort.

Studies of Adjuvant Therapy

In adjuvant studies of new anticancer drugs the usual efficacy parameters measured are survival and disease-free survival. A demonstration of improved disease-free survival without improved survival would not necessarily be sufficient for new anticancer drug approval, but it could be sufficient if it represented a real clinical gain, eg, if toxic therapy, pain, etc, were avoided by the delay. This must be resolved on a case-by-case basis.

SURVIVAL EVALUATION

Interpretation of survival data may be made difficult because the routine use of salvage therapies after failure of the test drug would seem to introduce an additional variable that could affect outcome. Since secondary treatments usually contribute little, if anything, to survival, however, it seems reasonable in most cases to use actual survival times in the analysis, not censoring at the time of initiation of salvage therapy. When the salvage regimen *is* highly active, results of the study can be seriously confounded. In this situation, unless all treatment arms receive the same salvage regimen, the contribution of the test treatment to survival may be difficult to assess. FDA must then decide whether to require additional studies of more appropriate design or to rely on other efficacy parameters as the basis for approval. Matters of this nature are best resolved between FDA and the new drug applicant *before* any studies are initiated.

An especially confounding situation in survival analysis of comparative studies occurs where there is crossover of treatments after treatment failure. Here, therapeutic activity of the salvage regimen is a real possibility. This situation should be avoided if at all possible, but avoiding it might prove difficult if both agents are very active. Time to progression of disease in that case might be the best parameter to assess. The cleanest situation, of course, is when failures on both arms of the comparative study are treated with the same secondary therapy.

Time to treatment failure and time to tumor progression are useful parameters, but FDA ordinarily does not accept them as surrogates for survival determination unless it has agreed in advance to do so for ethical reasons.

OBJECTIVE RESPONSE

The criteria for objective tumor response, including the criteria for progression, should be clearly defined. The objective response rate should be separately determined for all study patients, without any exclusion, and for the patients considered evaluable. If response rates differ by site of metastatic disease in a tumor type, separate rates should be determined for each site in addition to an overall response rate.

If lesions are initially small, the possibility of measurement error and also of diagnostic error is increased. The FDA believes cancer study protocols should require a minimum measurable lesion size for study entry. In comparative studies, readers of scans and x-rays should be blinded.

The submitted case report forms in a new drug application (NDA) should document objective response at each follow-up interval by recording the site of disease, diagnostic modality, and lesion measurements. All disease sites should be followed, and follow-up should not be limited to measurable disease sites, as progression may otherwise be missed. Follow-up intervals for all disease sites need not be the same. The information on individual case report forms is essential for proper study evaluation. It has been our experience that reported responses are often not documented on the case report forms and that in many instances the case report forms document that the reported response did not occur.

RESPONDERS VERSUS NONRESPONDERS

In reporting of single-arm uncontrolled studies with anticancer drugs, responders and nonresponders are often compared regarding survival. The FDA has not accepted a survival advantage in responders as evidence of drug benefit, because a capacity to respond (rather than the response itself) may be the cause of the better survival. In general the only way to assess the drug contribution is by comparison with an independent control group that did not receive the drug.

DRUG COMBINATIONS

Cancer chemotherapy often consists of drug combinations. If a new anticancer drug is to be used in combination with other anticancer drugs, we ask for clear evidence that it has an effect, ie, that the combination including the test drug be superior to the combination omitting the test drug. Of course we would really like to know the contribution of each drug in the combination, but we have not required this for NDA approval, concentrating instead on the particular drug in question. With respect to study design, when an anticancer drug is added to a combination and compared to the

combination without the addition of the test drug, analysis is straightforward. Substitution of the test drug for a drug in a combination and comparison with the original unsubstituted combination also present no problem with study analysis if an advantage for the new combination is demonstrated. But if the two combinations are equal in effect, it may be difficult to conclude that there was a drug effect by the test drug if the contribution of a drug it replaced in the original combination is not known.

END OF PHASE II MEETING WITH FDA

Drug firms are strongly encouraged to meet with FDA to discuss planned protocols prior to initiating the phase III comparative studies to be used for new drug approval.

REFERENCES

1. SACKS H, CHALMERS TC, and SMITH H. Randomized versus historical controls for clinical trials. Am J Med 72:233-240, 1982.
2. MOERTEL CG. Improving the efficiency of clinical trials: a medical perspective. Stat Med 3:455-465, 1984.
3. BYAR DP, SIMON RM, FRIEDWALD WT, ET AL. Randomized clinical trials. N Engl J Med 295:74-80, 1976.

DISCUSSION

Q: (Dr. Charles Moertel, Mayo) Dr. Johnson, perhaps I could get you to expand a bit on your orange juice and bubble gum theme and ask you about a positive decision which your group made regarding an NDA (new drug application) that I think has many of us in the oncology community very confused. I'm specifically referring to your decision which permitted marketing of the Infusaid implantable pump for hepatic arterial chemotherapy of colorectal cancer metastatic to liver. As you may know, such treatment places an extraordinary financial burden on either patient or third party with initial costs running up to $15,000 or more. This approach has achieved nationwide popularity and has been applied to thousands of patients in large measure because of Associated Press and United Press releases. These glowing statements announced that FUDR administered by the Infusaid pump method produced considerably fewer side effects than standard therapy for metastatic colorectal cancer (I presume 5-FU by weekly injection or the like). It was also said that the Infusaid pump would extend the life of patients with metastatic colon cancer up to 2 years. This particular press release was signed by Arthur Hull Hayes, Commissioner of the Food and Drug Administration. Try as I may, I have yet been unable to find one controlled study that substantiated Dr. Hayes' press release, much less two, and much less any study at all that gave convincing evidence of improvement in survival. As we and others have looked more closely at this method, we are becoming thoroughly convinced of the considerable hazard associated with it, including very serious and sometimes irreversible liver and biliary tract toxicity as well as gastrointestinal ulceration (which can be fatal). The evidence of contribution to survival is still lacking to the point that Dr. Wittes' program is currently sponsoring three randomized controlled studies to find out whether, in fact, Dr. Hayes' claim of improved survival has any truth to it whatsoever. Those of us who are conducting these studies are having a devil of a time getting patients enrolled because every small community hospital around the country is using this very expensive and hazardous therapy, and we can't get patients to come into a randomized trial. Could you explain this to me, please?

A: (Dr. Johnson) It's easy for me to explain because I wasn't involved. I agree 100%, 1000% with everything you've said . . . .

Q: (Dr. Moertel) And you intend to withdraw the NDA until appropriate documentation of therapeutic efficacy and safety becomes apparent?

A: I'm not sure we can withdraw the NDA, Dr. Moertel. Once we approve an NDA, it is very difficult to withdraw the approval. In order to withdraw it we have to prove the drug is an imminent hazard to the public health or that the drug is unsafe for its intended use or that on the basis of new evidence, not available to FDA at the time of approval, the drug is ineffective for its intended use. Before we approve a drug the burden of proof is on the drug company, but once we have approved it, that is reversed and it is very difficult to legally withdraw an NDA approval. Often the best we can do is ask the company if it will withdraw it voluntarily.

Q: (Dr. Moertel) Have they been asked?

A: I don't think they have.

Q: I think it probably would be possible to give a press release to United Press and Associated Press, retracting the first press release that came from the FDA. Would that not be possible, Dr. Johnson?

A: I'll promise you to look into it.

Design of Clinical Trials With Biological Response Modifiers

Ronald B. Herberman[1,*]

Biological response modifiers (BRMs) include biological and chemical agents which can increase host resistance against tumor growth and also biological agents which can have direct effects on tumor cells, by inducing cytolysis, growth inhibition, and/or differentiation. It is becoming increasingly clear that initial phase I clinical trials with BRMs need to be designed considerably differently from those for chemotherapeutic agents. In addition to determining the toxicity of each agent and its maximal tolerated dose, it is important to evaluate its effects on relevant immunologic and other host responses and to determine the optimal biological response modifying dose (OBRMD). Also, since most BRMs are likely to be effective mainly for treatment of cancer patients with low tumor burdens, and the biological response modifying effects of an agent may vary with the extent of disease, it seems necessary to first perform a phase IA trial in patients with advanced cancer, with an emphasis on determining toxicity and possibly the maximal tolerated dose. Then a phase IB trial will be performed, with patients with minimal or even undetectable tumor burden, to determine the OBRMD. These phase I trials will then allow planning for phase II trials for evaluation of antitumor effects, at doses and in cancer patients which might be expected to be favorable for detection of efficacy. [Cancer Treat Rep 69:1161–1164, 1985]

Biological response modifiers (BRMs) have been defined as those agents or approaches that modify the relationship between the tumor and host by modifying the host's biological response to tumor cells, with resultant therapeutic effects (1). BRMs may modify the host responses in several ways:

(*a*) by increasing the host's antitumor response by augmentation and/or restoration of effector mechanisms or mediators of the host's defenses, or decreasing a component(s) of the host's reactions which may interfere with antitumor responses;

(*b*) by administering natural or recombinant, genetically engineered effectors or mediators of antitumor responses;

(*c*) by augmenting the sensitivity of the host's tumor cells to endogenous mechanisms for control of tumor growth;

(*d*) by decreasing the transformation and/or increasing the differentiation of tumor cells; and

(*e*) by increasing the ability of the host to tolerate damage by conventional cancer treatments, particularly chemotherapy or radiotherapy.

A major category of BRMs includes the agents which would be called immunomodulators, and therapeutic approaches with such agents could be defined as immunotherapy. Immunomodulators would include both chemical and biological agents, which can modify immunological resistance mechanisms of the host that might be important in control of tumor growth or metastasis. There are a variety of different agents that are in this category. These include chemical or synthetic, as well as natural, agents that would affect the immune response, antigens on cells or isolated from them, immunologic effector cells or antibodies with antitumor effects, and soluble factors produced by various cells that could affect the immune system (referred to as cytokines). A type of nonimmunologic BRM which could favorably affect host resistance to tumors would be protective agents that in some way could overcome or lessen the damage which was produced by chemotherapy or radiotherapy and thereby facilitate the actions of the host's effector mechanisms.

Another category of BRMs would include cellular products, particularly products made by components of the immune system, that could have direct antitumor effects. For example, cytotoxic antibodies and some cytokines (eg, lymphotoxin and tumor necrosis factor) can have di-

[1]Biological Therapeutics Branch, Biological Response Modifiers Program, Division of Cancer Treatment, National Cancer Institute-Frederick Cancer Research Facility, MD.

[*]*Reprint requests to:* Ronald B. Herberman, MD, Pittsburgh Cancer Institute, 230 Lothrop St, Pittsburgh, PA 15213.

rect cytotoxic activity against tumor cells. Similarly, some immunologic effector cells, eg, cytotoxic T cells, natural killer (NK) cells, and macrophages, have substantial cytotoxic reactivity against a variety of tumor cells. There is also the potential for various cellular products to influence tumor growth, not by being directly cytotoxic, but by decreasing the state of transformation or increasing the differentiation of the tumor cells (optimally, to a terminal, nonproliferative mature form) or in some way changing the sensitivity of the tumor cells to control by intrinsic effector mechanisms in the host.

What are some of the problems related to rational design of trials with BRMs? A major problem is that most BRMs are very pleiotropic in their actions. It was hoped for some time that as we went from relatively crude microbial agents, for example, BCG or *Corynebacterium parvum*, to chemically defined, purified, even homogeneous substances, and particularly those with low molecular weight, only one function or action in the host would be observed. However, it has usually been found that low molecular weight homogeneous chemicals tend to be pleiotropic in their effects on the immune system or on other host functions. Another major problem which has been recognized recently is that most promising results obtained in preclinical studies, eg, marked immunomodulatory effects, have mainly been associated with administration of single doses of the agent. In contrast, for clinical therapeutic purposes, one generally wants to give the agent repeatedly over a prolonged period of time. Somewhat unexpectedly, it has been observed that upon repeated administration of various BRMs, hyporesponsiveness to at least some of the initial effects of the agent may develop. For example, as discussed elsewhere in this volume by Maluish and Herberman (2), repeated administration of alpha interferon to patients has been found to result in hyporesponsiveness to augmentation of NK activity. Another problem is that for some of the categories of BRMs, it is very difficult to design satisfactory or simple preclinical screening or evaluation procedures. Ideally, one would like to have all promising BRMs screened for their in vivo immunologic and antitumor effects in rodents. However, a variety of the potentially useful BRMs do not cross the species barrier. They might have potent effects, either directly on human tumor cells or indirectly on patients' effector mechanisms, but be entirely ineffective in rodents. For such BRMs, eg, most human interferons and monoclonal antibodies against human tumors, it has been necessary to rely on in vitro preclinical assessments and on studies of human tumor xenografts in nude mice.

As an example of some of the issues to be faced in studies with BRMs, I will now discuss some of the lessons learned from studies with the interferons. These represent the BRMs which have been most extensively evaluated over the last few years. The interferons represent a family of proteins which induce antiviral resistance in cells. With the advent of recombinant DNA technology, it has been possible to produce large, essentially unlimited quantities of homogeneous materials to work with. However, it has become very clear that each of the interferons that have been examined has quite pleiotropic effects. On one hand, they can directly inhibit the growth of certain tumor cells, and this could be an important mechanism by which they might be useful therapeutically. On the other hand, the interferons are very potent immunomodulators. They can augment NK cell activity and activate macrophages for antitumor cytotoxic activity. They can modulate specific T-cell immunity, and change various cell surface structures, eg, tumor-associated antigens and antigens of the major histocompatibility complex, that can be quite important in immune responses against tumors. Any one of these possible effects might account for therapeutic efficacy in some circumstances. For the last several years, the interferons have been in phase I and II clinical trials, and, by now, hundreds of patients have been treated. It has become reasonably clear that certain types of tumors, particularly lymphomas and hairy cell leukemia, are highly responsive to interferon. It is quite possible that the high degree of responsiveness which has been seen in these tumors may be related to a direct antitumor effect of interferon. There are at least some hints that higher doses are more effective than lower doses, but some excellent clinical responses have been observed with quite low doses. In other tumor types, in which clinical responses have been seen only infrequently, it seems possible that interferons have little direct antitumor effect but might be effective if they induced potent immunomodulation. As Maluish and Herberman (2) describe in more detail, the current clinical protocols that are being used with interferon have not been shown to have optimal or even good immunomodulatory effects on parameters which would be expected to be strongly affected, based either on in vitro studies with human peripheral blood mononuclear cells or on studies in animal model systems. A major concern is that one might draw a premature negative conclusion about the lack of efficacy of a BRM-like interferon when it has not been adequately tested under the appropriate circumstances. For evaluation of the potential of interferon to have a therapeutic benefit through some of its immunomodulatory effects, it would seem essential to develop protocols and schedules of administration that produce the desired alterations in the patients' immune functions. Until that has been achieved, there will not be an adequate test of the hypothesis that interferons can have therapeutic efficacy by their ability to augment immunologic effector mechanisms.

Another example of some current problems in the evaluation of BRMs has come from studies with another cytokine, interleukin 2 (IL 2). This is a product of T cells and large granular lymphocytes which can stimulate the growth and function of T cells and NK cells, and by either

one of these effects it might be quite useful therapeutically. The Biological Response Modifiers Program (BRMP) of the National Cancer Institute decided last year to become actively involved in clinical trials with IL 2, and, as a first step, it seemed important to evaluate it into our preclinical screening program. Preclinical evaluation of IL 2 has been easier than that of the interferons, since human IL 2 is active in mice and rats as well as in humans, so that it can be extensively tested both in vitro and in vivo in rodents as well as in humans. Detailed studies on how it might be effective in animal models are now being actively pursued. This should provide an extensive information base to help in the design of the clinical trials. In addition to the assessment of a variety of parameters in animal model systems, we have also been faced with the issue of deciding which IL 2 preparation to utilize. One question was whether to perform studies with natural IL 2, produced by lymphoid cells, or to proceed immediately to studies with the recombinant protein. On one hand, recombinant IL 2 was attractive because of the potential availability of unlimited quantities of homogeneous factor, whereas natural IL 2 preparations are difficult to make in large amounts and to purify to homogeneity. However, one might expect that the lack of glycosylation of the recombinant protein and possible differences in its physical form might make it appreciably different from natural IL 2 in terms of biodistribution or functional activities. Upon surveying possible supplies of IL 2, we soon became aware of 12 different products, six natural and six recombinant. Since clearly it would not be practical for us to perform clinical trials with more than one or two of these preparations, a series of preclinical studies were performed to directly compare the potency of the different products in various functional assays and to examine them for possible contaminants or unexpected effects (3).

To adequately evaluate the relative potency of the various IL 2 preparations, we first had to standardize them in terms of their activity in assays of promotion of growth of T-cell lines. When the preparations were all compared to a BRMP standard preparation of IL 2, it was obvious that a unit of IL 2 from one company was considerably different from a unit of IL 2 from another company. Overall, there was a > 100-fold variation in the preparations in regard to the relationship between the companies' stated units of activity and the BRMP reference units which were actually measured. To adequately compare the results of various trials with different IL 2 preparations, it will obviously be important to achieve standardization of the units of activity. Beyond the standardization of the potency, another issue was to determine the possible presence of contaminants in the preparations. Although it is desirable to work only with homogeneous materials, there were no completely homogeneous natural IL 2 preparations available, and it was therefore necessary to screen for a number of other cytokine activities and other materials which might be contaminating the IL 2. Endotoxin was detected in some of the natural and recombinant preparations, which would clearly be undesirable for clinical use. Some of the natural products were found to have interferon activity, and if these were used in clinical trials it would be difficult to determine whether any therapeutic effects were due to the IL 2 or to the interferon. In addition, after the IL 2 preparations were standardized to have the same potency in terms of their ability to grow T cells, they were still found to vary in regard to other biological activities. For example, there was much heterogeneity, > 50-fold, in the ability of the preparations to augment NK cell activity, another important property of IL 2. Even among the recombinant IL 2 preparations, there was a 50-fold difference among the preparations in NK-augmenting activity. Another unexpected result from this comparative study was the observation that most IL 2 preparations, including the recombinant materials, had macrophage- or monocyte-activating properties. This indicates an additional functional effect of IL 2 which might have therapeutic implications and will need to be monitored during clinical trials with IL 2.

Another major issue to be considered is the design of phase I clinical trials with BRMs. Currently, clinical trials with BRMs are being performed along lines very similar to those utilized for trials with chemotherapeutic agents. Phase I trials with both BRMs and chemotherapeutic agents are performed with patients with advanced cancer who have already failed other therapies. The major emphasis on the phase I trials has been to evaluate toxicity and to determine the maximum tolerated dose (MTD). To get some information about the immunologic or other biologic effects of the BRMs, assays have been performed to determine possible biological responses in the context of the dose escalation studies. A major goal in the phase I trials with BRMs is to not only determine the MTD but also to determine the optimal biological response modifying dose (OBRMD), namely, the dose that would have the optimal or maximum effects on the immunologic or other biologic parameters that would be expected to be most important for therapeutic efficacy. Unfortunately, with current trial design, one regularly gets the information about toxicity and MTD, but often inadequate information is gathered to determine the OBRMD. For example, in the studies with interferon, although hundreds of patients now have been put into phase I and II trials, the OBRMD has yet to be determined. This has resulted in the design of phase II trials with doses close to the MTD. Since the interferons might have therapeutic effects as a result of their ability to modify biological responses, it would be highly desirable to also perform phase II therapeutic trials with the OBRMD, which might be considerably different from the MTD. Unfortunately, it is becoming increasingly clear that a phase I trial that is designed primarily to obtain MTD and toxic-

ity information may not be adequate to get sufficient information about the OBRMD.

Another major problem to consider in this regard is that the effects of BRMs on the immune system or on other biological responses, in previously treated patients with advanced disease and large tumor burdens, might not be comparable to the effects of the same materials, even at the same doses, on patients with lower tumor burdens. There is also considerable evidence from animal model systems that many, if not most, BRMs would be most likely to only have therapeutic efficacy in individuals with low tumor burdens, and to be ineffective in individuals with extensive, widespread disease. Therefore, with the current design of phase I and II clinical trials with BRMs, immunomodulatory and positive clinical effects of potentially valuable BRMs might be completely missed.

Yet another problem is one of adequate systemic delivery of the agents to the tumor sites. A number of BRMs are labile, have a short half-life in the circulation, and do not distribute evenly throughout the body. IL 2 is an example of this problem. After iv inoculation, its half-life in the serum is only a few minutes, and it may not accumulate in sufficient levels in the lymphoid organs or at sites of tumor growth, where one would hope to have it mediate its effects.

What are the possible solutions to the problems that I have raised? One would be to have a greater emphasis in the phase I trials on determining the OBRMD, defined as the safe route, dose, and schedule which induce sustained or repeated alterations in a maximum number of the desired parameters, particularly immunologic functions that would be expected to have effector activity directly against the tumor. One could also examine the biological response modifying aspects in phase I trials in patients with lower tumor burden, and even in patients with minimal or no detectable disease who have not received chemotherapy or radiotherapy. Alternatively, or in addition, one could perform phase I trials with BRMs in combination with cytoreductive therapy, since these combinations may be most useful for therapy and it would be important to first determine the biological response modification which could be induced in this context.

The sequence of clinical studies with BRMs which I have proposed and which has been supported quite enthusiastically by the BRMP's Decision Network Committee would be to divide phase I into two stages, phase IA and IB. Phase IA trials would first be performed in patients with advanced cancer, primarily to get the necessary information on toxicity and possibly to determine the MTD. Immunologic and other biologic studies would be performed in those patients, but as in current studies, this would be a secondary objective. At the conclusion of the phase IA trial, a phase IB trial would be performed, in which the main focus would be on determining the OBRMD. Such studies would be performed in patients with minimal, or perhaps even undetectable, tumor burden. Then the established OBRMD would be used to examine therapeutic efficacy in phase II trials in patients at the same stage of disease as the patients utilized in the phase IB trial.

This is clearly a more complicated study design than that currently used. Only certain types of patients would be suitable for the phase IB or phase II trials that I am proposing. However, unfortunately, there are several major tumor types in which after surgical resection of tumor, there is a relatively low tumor burden, but a high probability of rapid recurrence of tumor growth and metastases and refractoriness to currently available forms of therapy. For example, patients with non-small cell carcinoma of the lung or Dukes' C colorectal carcinoma would seem to be good candidates for phase IB and II studies with BRMs.

BRMs appear to have considerable promise for therapy of malignant disease. This conclusion is based mainly on many positive therapeutic results with a wide variety of BRMs in experimental animal tumor systems (4). The major challenge appears to be how to translate such therapeutic efficacy into a clinically applicable setting. As discussed above, it seems that empirically designed clinical studies, even with highly promising BRMs, will be unlikely to produce even close to optimal therapeutic efficacy. Rather, it will probably be essential to systemically evaluate the range of biological effects of each agent, determine the OBRMD in well-designed phase IA and IB studies, and only then perform the appropriate phase II and III therapeutic trials.

REFERENCES

1. MIHICH E, and FEFER A, eds. Biological response modifiers: subcommittee report. Natl Cancer Inst Monogr 63:1–252, 1983.
2. MALUISH AD, and HERBERMAN RB. Assessment of biological responses: what to measure and when. Cancer Treat Rep 69:1165–1169, 1985.
3. THURMAN GB, MALUISH AE, ROSSIO JL, ET AL. Comparative evaluation of multiple lymphoid and recombinant human IL-2 preparations. *In* Proceedings of the 4th Lymphokine Workshop, in press.
4. TALMADGE JE, FIDLER, IJ, and OLDHAM RK. Screening for Biological Response Modifiers: Methods and Rationale. Boston, Martinus Nijhoff Publishing, 1985.

Assessment of Biological Responses: What To Measure and When[1,2]

Annette E. Maluish* and Ronald B. Herberman[3]

During the past several years, increasing attention has been devoted to the use of biological response modifiers (BRMs) for the treatment of cancer. Phase I trials of BRMs must be designed to provide information not only regarding toxic effects, pharmacokinetics, and antitumor activities, but also on the many immunomodulatory effects. Biological monitoring must be carefully planned to enable meaningful conclusions to be drawn as to the optimal dose, schedule, and route of administration. Important considerations include selection of assays for the various biological effects, timing of testing following BRM administration, and methods for data interpretations. [Cancer Treat Rep 69:1165–1169, 1985]

During the past several years, increasing attention has been devoted to the potential application of biological response modifiers (BRMs) for the treatment of cancer. It has been suggested that BRMs might provide a novel approach to therapy and be an effective and less toxic alternative or adjunct to conventional forms of treatment such as chemotherapy, radiotherapy, and surgery. BRMs have been defined as "those agents or approaches that modify the relationship between tumor and host by modifying the host's biological response to tumor cells with resultant therapeutic effects (1)."

BRMs mediate their effect in a variety of ways. These include activating, increasing, or restoring reactivity of immunological effector mechanisms that are involved in resistance to tumor growth and metastasis or inhibiting suppressor mechanisms which interfere with effective host resistance to tumors. Some BRMs may have direct antitumor effects in addition to their immunomodulatory properties. The rationale for using BRMs for cancer therapy is that by restoring deficiencies in immunological function and/or augmenting effector function above the existing levels, the host's biological response will overcome the tumor growth. To have a high likelihood to achieve such an effect, clinical trials must be designed with the specific stated aim of determining the optimum conditions under which to administer the BRM. It is not an easy task since there are many questions that need to be addressed. These include such issues as the optimal dose, frequency, and route of administration. In developing strategies for laboratory monitoring of patients to determine an optimum BRM dose and schedule, the obvious question is "which biological response does one wish to modify?" This presumably will vary with the BRM, but in general it is necessary to measure several functions serially in patients treated with a variety of schedules and doses of an agent.

The use of the monitoring data obtained in phase I studies for the design of phase II and III studies is discussed elsewhere in this volume by Pinsky (2). In this presentation, we will discuss the general principles for the selection of assays and the times for testing during phase I studies with BRMs.

It should be noted that the emphasis here is on the strategy for phase I testing with BRMs, with the objective of determining an optimum BRM dose and schedule. Another important aspect of monitoring is to determine the possible correlation between immunological effects and antitumor response or to determine the predictive value of a certain test. However, this objective is better carried out in phase II studies, which are designed to determine therapeutic efficacy and usually are performed with more defined patient populations.

SELECTION OF ASSAYS

Most BRMs, even genetically engineered homogeneous

[1]Supported in part by contract N01-CO-23910 from the Office of the Director, National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

[2]The views expressed do not necessarily reflect those of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

[3]Clinical Immunology Services, Program Resources, Inc, and Biological Therapeutics Branch, Biological Response Modifiers Program, National Cancer Institute-Frederick Cancer Research Facility, MD.

*_Reprint requests to:_ Dr. Annette E. Maluish, Clinical Immunology Services, National Cancer Institute-Frederick Cancer Research Facility, PO Box B, Frederick, MD 21701.

materials, are pleiotropic in their actions. It is necessary, therefore, in selecting functions to be monitored, to cover as many of these actions as possible. The selection of assays is made with the following general considerations.

Assays for Effector Functions

Assays for effector functions might be most important, since activated effector cells might account for, or contribute to, the antitumor effects of the BRM. Historically, much emphasis has been placed on cytotoxic T cells, and these can clearly be shown to play an important role in defense against immunogenic tumors expressing tumor-associated transplantation antigens. In the past several years, natural killer (NK) cells have been shown to contribute to host resistance against a variety of tumors and appear to be effective in prevention of metastatic spread (3). In experimental systems, macrophages have also been shown to be effective against growth of metastases (4). NK cells, macrophages, and even granulocytes are clearly important immune effector mechanisms and function as part of the natural defense system. These functions can be activated by a number of different BRMs and can react against a wide variety of tumor cells, including those without detectable tumor-associated surface antigens (3,5).

Ideally, assays for each of these effector functions, especially of cytotoxic T cells, NK cells, *and* monocytes, would be carried out, but due to the numbers of cells needed for each assay, the limitations on the quantity of blood that is available in clinical studies, and the usual need to monitor other functions, the choice of which effector assay(s) to utilize must be made after consideration of the known actions of the BRM under study. For example, in a study of the effects of interleukin 2 (IL-2), all three effector functions need to be measured, since IL-2 has been found to affect all three functions.[4]

Assays for General Immune Competence

Lymphocyte proliferation in response to mitogens and/or mixed allogeneic leukocytes has been used for many years as a measure of immune competence, particularly of T-cell function. Since cancer patients, especially those with advanced disease, often have depressed immunocompetence and many BRMs can stimulate or restore T-cell functions, this assay is usually included in the panel of assays in addition to assays for effector functions.

With the advent of monoclonal antibodies to a variety of cell surface markers and highly sophisticated analysis using flow cytometry, a detailed and quantitative evaluation of possible changes in cell subpopulations can be performed, which may provide important indications of shifts in the proportions of various cell types or changes in levels of activation or differentiation. This analysis should involve an evaluation of the density of expression of the markers as well as the percentage of cells expressing the marker. For example, 90%–95% of monocytes are positive for expression of HLA-DR, but the level of expression may be relatively low. On activation with IFN gamma, the percent of positive monocytes does not change, but the level of expression increases dramatically in a dose-dependent fashion. In general, the panel of antibodies should include anti-T3 (Leu 4; total T cells); anti-T4 (Leu 3; T-helper/inducer cells); anti-T8 (Leu 2; suppressor/cytotoxic cells); anti-Leu 12 (B cells); anti-Leu 11, which reacts with most large granular lymphocytes and therefore is very useful to enumerate the NK cell population; and anti-Leu M3 (monocytes). Evaluation of other markers for activation, eg, HLA-DR, transferrin receptor, and IL-2 receptor, should be added for relevant BRMs.

Activities Shown To Be Altered by the BRM In Vitro

A survey of the effects of a BRM on normal human cells in vitro can provide very useful guidance for selection of the relevant assays for inclusion in the panel for monitoring. Unless the BRM cannot act directly and requires in vivo activation or accessory factors, its direct in vitro effects should serve as a reliable guide to its potential biological response modifications.

In addition to effects on the immunologic cells, many BRMs also have other biological response modifying effects, eg, alterations in the production of lymphokines and other cytokines and production of various cellular growth factors, such as colony-stimulating factors (6). Many of these factors can be assayed in serum or plasma and since many are regulatory molecules, an assessment of their production provides additional information on the activity of the BRM. Examples of such regulatory molecules include interleukin 1 (IL-1) and interferon gamma, which should be assessed following treatment with IL-2.

In response to some BRMs, the levels of certain enzymes increase within the cell. The 2′5′-oligo A synthetase (2′5′A) and protein kinase are produced in lymphoid cells. The detection of these enzymes in the cell is a clear indication of exposure to interferon. An assay such as this was very informative in a recent study using the interferon-inducer polyinosinic-polycytidilic acid, which was stabilized with poly-L-lysine and carboxymethyl-cellulose (Poly IC-LC). In this study, Poly IC-LC was given at low doses (1 mg/m^2 im) to cancer patients and, although no detectable levels of IFN were seen in the serum, the elevated levels of 2′5′A in the cells were a clear indication of the production and effects of IFN (7).

[4]Thurman GB, Maluish AE, Rossio JL, et al. Comparative evaluation of multiple lymphoid and recombinant human IL-2 preparations. Manuscript submitted for publication.

Functions Shown To Be Altered in Preclinical Studies

With a wide variety of potential assays to be performed for each BRM, the decision of which functions should be emphasized can be aided by evaluation of the agents in animal models, to determine which cell populations or functional changes are required for the antitumor effects of the BRM. It is possible, for example, to selectively deplete mice of certain cell types and evaluate the agents in these animals. For example, nude mice can be used to evaluate the efficacy of a BRM in T-cell-deficient individuals. Similarly, beige mice can serve as a strain of mice with a selective deficit in NK activity. Alternatively, mice can be selectively depleted of NK cells by administration of antibodies selective for NK cells (eg, antiasialo GM_1).

Data from the Biological Response Modifier Program's Preclinical Screening Laboratory are very useful as guidelines for predicting the timing of effects on various functions. Animals can be treated with a BRM at varying intervals and assays performed to determine the optimum schedule for immunomodulating or other relevant effects. Caution must be used, however, in the interpretation of these data, since the level of activity in one organ (eg, spleen) may be quite different from that in the peritoneum or peripheral blood. Species restrictions in the activities of some BRMs also may limit this approach. For example, some human cytokines (eg, IFN) may be ineffective on rodent cells. Similarly, most monoclonal antibodies to human tumor-associated antigens will not have sufficient cross-reactivity with mouse cells to allow a complete assessment of their effects on host functions. However, even in this case, studies of the effects of monoclonal antibodies on human tumors growing in nude mice can be useful, not only for assessing direct antitumor effects, but also for determining the possible involvement of antibody-dependent cellular cytotoxicity, since some monoclonal antibodies can interact with the Fc receptors of mouse effector cells.

Assays for Direct Effects on Tumors

Some BRMs (eg, IFN) are known to be directly antiproliferative to a variety of tumor targets. Assays measuring the antiproliferative effects of sera against both lymphoid and solid tumor cell lines provide information on the possible production of antiproliferative factors. In a study of interferon alpha in our laboratory, there is evidence for the induction of new antiproliferative factors in the sera of some patients treated with interferon alpha.[5]

[5]Imai M, Maluish AE, and Herberman RB. Anti-proliferative activity in sera of patients receiving interferon. Manuscript submitted for publication.

Some BRMs (eg, interferon and/or retinoic acid) may also exert effects on the state of differentiation of tumor cells, and induction of terminal differentiation has important therapeutic implications. For studies in patients with leukemia, such effects can be readily monitored on the circulating tumor cells. For other tumor types, it would be necessary to perform serial biopsies of tumors, to determine whether administration of the BRM had induced a change in the level of differentiation.

How Many Assays To Perform

Since a number of BRMs not only augment the activity of some particular functions but can also cause depression of other functions, it would be ideal to measure as many functions as possible. Because of limitations in the number of assays which might be performed on the blood specimen from a given patient, one strategy might be to measure a different subset of functions in each patient and then pool the results to obtain an overall pattern of biological effects. However, the initial studies of new BRMs are usually performed on patients with a variety of malignancies and in various stages of disease. Since the effects observed are not always identical from patient to patient, and if any correlations between assays or cause-and-effect relationships are to be postulated, the activities need to be measured in the same specimens.

Thus, the possible assays need to be prioritized and this priority will vary somewhat from one BRM to another. In general, the assays for effector function, enumeration of the major subpopulations of cells (but not all surface markers), and the assays for functions known to be affected in vitro would be given the highest priority. The miniaturization of many assays such as assays for monocyte-mediated growth inhibition of tumor target cells and for lymphocyte proliferation has reduced somewhat the need to make difficult choices and enables more assays to be performed on each blood sample.

The timing of the assays is dependent to a certain extent on the particular BRM and schedule of administration and, of course, one needs to consider the total amount of blood which can be obtained safely during the course of the study (usually a maximum of 500 ml). In our experience using miniaturized assays, the panel of selected assays can be performed on 40 ml of blood. It is important to draw blood at the same time each day (preferably morning) since a circadian rhythm has been described for a number of functions. To assess changes in function, determinations at 24 and 72 hours following a single dose of most BRMs should be adequate. In addition, it is imperative that several baseline samples are obtained (*see* below). We monitored a group of patients receiving Poly IC-LC at 2, 6, and 24 hours following administration and observed almost complete disappearance of NK activity at 2 and 6 hours. This was attributable to a disappearance from the peripheral blood of the effector cell popula-

tion, as detected by flow cytometry studies. It is conceivable that some cell subpopulations may be induced to leave the circulation soon after administration of BRMs and that flow cytometry studies can detect such shifts and provide explanations for transient deficits in functional activities. Re-equilibration of cell distributions usually occurs within 24 hours and, therefore, for the evaluation of lasting functional change, measurement at 24 and 72 hours would seem more appropriate than immediately following treatment.

Interpretation of Change

A major problem in the analysis of clinical monitoring data is the substantial spontaneous fluctuation in reactivity in the assays over a period of time. These fluctuations may be attributed partially to technical factors in the assays, eg, changing susceptibility of target cells to lysis. They may also be caused by often considerable biological variability.

The problem of assay variability can be reduced by either testing all samples from each patient in a single assay or by testing in different assays but including controls which enable the "normalization" of the data. Testing of all samples in a single assay can be accomplished by cryopreservation, by controlled-rate freezing of each cell preparation within a few hours after collection. This is a valuable approach for assessment of many activities. Clearly, the function must be known to be reliably cryopreserved. For monocyte-mediated cytotoxicity and cytostasis, enumeration of lymphoid cell subpopulations, and assessment of lymphocyte proliferation responses, this is a useful technique (8). Unfortunately, NK activity cannot be reliably cryopreserved on each sample. This appears to be due to variable detrimental effects on cytotoxic function, since there is not a selective loss of functional large granular lymphocytes and since the deficit can be restored following incubation of cryopreserved cells at 37°C for 5–12 hours. However, we do not as yet feel that this is reliable and cannot recommend the procedure for monitoring of NK function. It is therefore necessary to test each sample immediately after collection and procedures need to be devised to control for day-to-day technical variations. This is accomplished by collection of large numbers of cells from normal donors by leukapheresis and cryopreservation. The recovery of NK activity is consistent in multiple aliquots of a single sample and remains quite high in some specimens for normal donors. Specimens with high activity are therefore chosen as controls for this assay. Three such normal cryopreserved controls are included in every assay and are used to normalize the data obtained on each patient (9).

The other cause of variability is the intrinsic fluctuations in immunological reactivity of each donor. These fluctuations are more pronounced for some functions than others. Lymphocyte proliferative responses to mitogens have been found to be quite stable over time, whereas NK activity is rather variable. It cannot be assumed that variability determined in normal donors is necessarily the same in cancer patients, and each function must be evaluated in both normal donors and untreated cancer patients. There may be quite substantial differences between normal donors and patients with advanced cancer. It is especially difficult to distinguish between treatment-induced changes in reactivity and a trend for initially high or low values to spontaneously shift toward the normal range (statistical principle of regression towards the mean). It is useful to obtain several (at least two or preferably three) repeated baseline specimens from patients prior to BRM treatment. This enables a more accurate assessment of the mean baseline value. The limits of variability can be calculated from repeat testing of normal donors and cancer patients and need to be established for each assay being used. For example, these limits of variability for NK activity have been found to be 15% cytotoxic above and below the mean of the pretreatment values for a patient. That would mean that for a patient with a mean pretreatment value of 30% cytotoxicity, the limits of variability would be set at 45% and 15% cytotoxicity. In other words, if that patient was not treated, his NK activity (after correction for assay variation) might fluctuate in a random fashion from 15% to 45% cytotoxicity. Only changes above or below these limits can truly be considered as treatment-induced changes. We found these limits to be considerably different for different assays (eg, ± 10% for monocyte-induced growth inhibition of tumor targets). This is an important concept, and considerations of both assay and biological variations are essential if true treatment-induced changes are to be identified.

CONCLUSION

A primary concern in a phase I trial with a BRM is to carefully monitor for the biological response-modifying effects of the agent at various doses and schedules of administration. The trial must be designed with this aim as a major criterion in deciding on doses, schedules, and routes of administration, giving full consideration to factors such as half-life of the agent. If, and only if, the numerous aspects of the biological monitoring are considered and carefully planned, then meaningful conclusions can be drawn and progress made toward using BRMs to optimally augment the host's biological responses. The determination of whether the optimum BRM dose is also the optimum therapeutic dose must await the conclusion of more trials specifically designed to test this question.

REFERENCES

1. MIHICH E, and FEFER A, eds. Biological response modifiers: subcommittee report. Natl Cancer Inst Monogr 63:251, 1983.

2. PINSKY CM. Applicability of phase I trial results in the design of phase II and III biological response modifier trials. Cancer Treat Rep 69:1171-1173, 1985.
3. HERBERMAN RB, ed. NK Cells and Other Natural Effector Cells. New York, Academic Press, 1982, p 1566.
4. FOGLER WE, and FIDLER IJ. Role of macrophages in host resistance against tumors. *In* Basic and Clinical Tumor Immunology (Herberman RB, ed). The Hague, Martinus Nijhoff, 1985, in press.
5. HERBERMAN RB, and ORTALDO JR. NK cells and natural defenses against cancer and microbial diseases. Science 214:24-30, 1981.
6. SCHLICK E, BARTOCCI A, and CHIRIGOS MA. Effect of azimexone on the bone marrow of normal and γ-irradiated mice. J Biol Response Mod 1:179-186, 1982.
7. MALUISH AE, REID J, LEVY H, ET AL. Immunomodulatory effects of poly ICLC in cancer patients. J Biol Response Mod, 1985. In press.
8. MALUISH AE, ORTALDO JR, SHERWIN SA, ET AL. Changes in immune function in patients receiving natural leukocyte interferon. J Biol Response Mod 2:418-427, 1983.
9. MALUISH AE, ORTALDO JR, CONLON JC, ET AL. Depression of natural killer cytotoxicity after in vivo administration of recombinant leukocyte interferon. J Immunol 131:503-507, 1983.

Applicability of Phase I Trial Results in the Design of Phase II and III Biological Response Modifier Trials

Carl M. Pinsky[1,*]

The topic of this presentation is the applicability of phase I trial results in the design of phase II and III trials of biological response modifiers (BRMs). Clinical trials using BRMs have been given separate attention in this volume because in many ways trials with BRMs are different in their design and implementation from trials with chemotherapeutic drugs. Drs. Herberman and Maluish have already described some of these characteristics so I will repeat points that they made only for emphasis, or as a means of describing the thought process that one must go through in designing phase II and III trials with such agents. In discussing the applicability of the results in phase I trials to the planning of phase II and III, one should look at the different implications of the word "applicability." First, we must ask ourselves, "Are the results of phase I trials applicable at all to the design of phase II and III trials with BRMs?" If the answer to that question is yes, then we must discuss the techniques we use to apply these results in the design of the latter trials. On the other hand, if the answer is no, the results are not applicable, then we must go back to our phase I trials and redesign them so that the results will be applicable.

What is different about BRM trials that sets them apart from trials with standard cytotoxic chemotherapeutic agents? Table 1 lists the hypotheses that fundamentally underlie most BRM trials. First, it is believed that the development, existence, and progression of cancer are evidence that there must be some defect in the biological responsiveness of the patient who has the cancer. This is analogous to at least one hypothesis underlying many chemotherapeutic trials: namely, that cancer cells fundamentally differ in their biology and biochemistry from normal cells. Both hypotheses lead to the concept that the differences, either in the patient or in the tumor cell, are exploitable in a therapeutic sense to turn the advantage in favor of the patient over his cancer. It is worth mentioning that there is a great deal of evidence to suggest that this hypothesis is true and that defects in various host responses are frequent in patients with cancer (1-4). Some studies indicate that similarly staged patients have a poorer prognosis when immunodeficiency is also present (2,3). What is not at all clear is whether or not the deficiency preceded, and can be one of the causes for, the cancer as in primary (1) or iatrogenic (4) immunodeficiency, or whether the deficiency appears only as the cancer progresses. At any rate, the data from the first hypothesis lead naturally to the second hypothesis, which says that if you correct the defect you can expect that the tumor will regress. This may or may not be correct for most tumors, but there surely is good evidence that at least under certain specific instances this can be the case. Probably, the best documented instance is when a cancer is transplanted with another organ, such as a kidney. To allow the recipient to retain the organ transplant, immunosuppressive drugs are often required. Cancers which arose de novo, or which were accidentally transplanted in this instance, became widespread (4-6). In several cases, the only therapeutic maneuver which was necesary to cause regression of these tumors was removal of the immunosuppression by stopping the immunosuppressive drugs (4-6). This corrected the defect and led to regression of the cancer. If similar immune defects could be detectable and corrected in patients with other cancers, then similar regressions of tumor might be expected. This then leads to the third hypothesis, which says that BRMs are capable of correcting the defect. This hypothesis is probably the most difficult to prove correct. Actually, most trials of immunotherapy in years past, or biological response modification in more recent years, have not consistently yielded data which would suggest that the host-response defects of cancer patients were easily correctable. On the other hand, and this goes right to the heart of the presentations in this volume, the study design of most trials with BRMs has not adequately addressed this question. The problem is that BRM trials looking for correction of host defects usually have not been carried out with therapeutic intent, and most of the BRM trials carried out with therapeutic intent have not adequately measured host responses. Bringing these two aspects together, either in a single trial or in a series of

[1]Memorial Sloan-Kettering Cancer Center and Cornell University Medical College, New York, NY.

[*] *Reprint requests to:* Carl M. Pinsky, MD, Biological Response Modifiers Program, National Cancer Institute, Frederick Cancer Research Facility, Bldg 426, Rm 1, Frederick, MD 21701.

TABLE 1.—Hypotheses to be tested in clinical trials of BRMs

1. Presence of cancer implies a defect in the host.
2. Correction of the defect will cause regression of the cancer.
3. BRMs can be given in a manner which is capable of correcting the defect.
4. Administration of BRMs in a way which corrects the defect will induce tumor regression.

TABLE 2.—Aims of phase I trials

1. Determination of maximum tolerated dose
2. Selection of dose, route, and schedule for phase II
3. Evaluation of clinical pharmacology

well-designed and -executed trials, is the challenge we now face. If this is not done, then we will have never tested the hypotheses that we claim we are trying to prove or disprove. In all fairness, it should be noted that a number of investigators have made attempts to determine the optimal regimen for biological response modification in phase I trials (7-10). What has not been accomplished, so far, is the definition of a clear-cut optimal program to carry to phase II. This has precluded the testing of the fourth hypothesis, which states that if BRMs are used appropriately to correct the defect(s), then the cancer will regress. The fourth hypothesis would be tested in phase II and III trials, while the third hypothesis has to be tested in phase I.

In table 2, the aims of all phase I trials of biological as well as chemotherapeutic agents, are listed. The first aim is to determine the maximum tolerated dose. This includes gathering data about expected and unexpected toxicity. Due to the limited number of patients in phase I trials, it is unreasonable to expect to get a true picture of the range of toxic effects which may be seen in the general use of an agent before phase II and III trials are complete. The second aim is to select a dose, route of administration, and schedule that will be carried to phase II trials for estimating efficacy. The third aim of phase I trials is to obtain data relating to the clinical pharmacology of the agent in question. In the case of BRMs, it would seem that the maximum tolerated dose is not necessarily the dose that might be optimal for correcting the defect in biological responses, and instead we have evolved, as Drs. Herberman and Maluish have already described, an optimal dose and schedule for biological response modification. Detecting that regimen is the purpose of phase I trials with BRMs and using that information in the appropriate design of phase II is the topic of my discussion. The evaluation of clinical pharmacology of an agent should include not only measuring half-life in the blood or other tissues (after various doses, routes of administration, and schedules), but also obtaining data on biological response modification which are appropriate in testing the hypotheses that we have already described. As such, we would anticipate that well-designed phase I trials of BRMs will very quickly establish maximum tolerated dose (to give an indication of the range in which we can work) and will then be directed toward the establishment of an optimal dose and schedule for biological response modification. When we go to plan the phase II and III trials, we will therefore be faced with the analysis of a great deal of data relating to dose given during phase I and resultant biological responses. It is probably appropriate to point out at this time that in many trials which have already been completed with biologics the question is not how high one went, but rather how much lower could the dose have been and still yielded biological responses?

In the remainder of the discussion, I will assume that the essentials of phase I trials have already been satisfied and that adequate numbers of patients have been treated at various dose levels and schedules. What decision rules should be used in planning for the phase II trials? As has already become apparent from the presentation of Dr. Maluish, when various dose levels are used and numerous laboratory assays are obtained at each dose level, it is conceivable that the optimal dose causing favorable changes in one assay may not necessarily be the optimal dose causing favorable changes in another. I would propose in such a case that if the doses causing the appropriate response are relatively close together a compromise dose should be chosen. In the case where the optimal doses in different assays are widely separated, it might be reasonable to do a two-dose phase II trial; ie, treat one cohort with the lower dose and another cohort with the higher dose. In fact, a perfectly acceptable approach would be to use two cohorts and a crossover design. Such a design is also appropriate for evaluating two or more schedules or for comparing maximum tolerated dose with optimal dose for biological response modification.

The foregoing discussion has concerned the use of biologics as single agents in phase I and II trials. Appropriate phase III trials using BRMs as single agents would follow any phase II trials in which biological response modification and antitumor efficacy could be established. On the other hand, it is already clear, particularly from experimental animal model systems, that some BRMs might best be used as adjuvants to other standard therapies, in particular, in combination with cytotoxic chemotherapeutic agents. Already in the literature of clinical immunotherapy, a number of agents have been shown to have antitumor activity when used with chemotherapy in phase III randomized trials. Still, one has a gnawing sense of unease that the dose, route, and schedule employed have never been established by appropriate phase I trials. In fact, one suspects that even better results might have been seen had these trials been preceded by phase I trials in which the optimal changes in immune responsiveness were evaluated using the agent combined with appropriate chemotherapy.

It is not easy to envision the design of such a phase I trial. The obvious solution would seem to be to first use

an established chemotherapy program and to perform a pilot trial in which various dose and schedule manipulations of the BRM were superimposed, with the endpoint being modifications in biological responses which were considered to be consistent with the expected mechanism of action of the experimental agent. Such a design would have an additional advantage, namely, it would satisfy the ethical concerns of many people that patients with cancer had received appropriate chemotherapy during the trial. I believe that it is fair to say that few, if any, phase I trials with combinations of chemotherapy and BRMs have been carried out to date. One of the challenges we face in this field is to imaginatively approach this question. If the phase I trials are carried out in a manner which is clearly goal-directed toward planning a combination which will be tested in phase II and eventually in phase III, then one satisfies the applicability question right up front. Similar considerations can be invoked for combination therapies with surgery and/or radiation therapy.

In conclusion, I have suggested in this presentation that most immunotherapy and BRM trials in the past have not addressed the fundamental hypotheses which should underlie these approaches. On the other hand, more recent work, at least in phase I, has approached the question of the dose of the BRM as it relates to the endpoint of biological response modification. It is now time for phase II trials to begin which utilize this information in an appropriate manner. Innovative study designs will be required to deal with the complexity of the resultant data sets. Once these are completed, and hopefully some agents show biological response modification and antitumor efficacy, phase III designs will be possible. In addition, a more imaginative approach to the initial design of therapies using combinations of standard therapies with BRMs is urgently needed.

REFERENCES

1. GOOD RA, and FINSTAD J. Essential relationship between the lymphoid system, immunity, and malignancy. Natl Cancer Inst Monogr 31:41-58, 1969.
2. EILBER FR, and MORTON DL. Impaired immunologic reactivity and recurrence following cancer surgery. Cancer 25:362-367, 1970.
3. PINSKY CM, EL DOMEIRI A, CARON AS, ET AL. Delayed-hypersensitivity reactions in patients with cancer. Recent Results Cancer Res 47:37-41, 1974.
4. STARZL TE, PORTER KA, ANDRES G, ET AL. Long-term survival after renal transplantation in humans: (with special reference to histocompatibility matching, thymectomy, homograft glomerulonephritis, heterologous ALG, and recipient malignancy). Ann Surg 172:437-470, 1970.
5. PENN I. The occurrence of cancer in immune deficiencies. Curr Probl Cancer 6:1-64, 1982.
6. WILSON RE, HAGER EB, HAMPERS CL, ET AL. Immunologic rejection of human cancer transplanted with a renal allograft. N Engl J Med 278:479-483, 1968.
7. HIRSHAUT Y, PINSKY CM, and OETTGEN HF. Design of phase-I trials of immunopotentiators for cancer therapy: levamisole and *Corynebacterium parvum*. Ann NY Acad Sci 277:252-259, 1976.
8. DILLMAN RO, BEAUREGARD JC, MENDELSOHN J, ET AL. Phase I trials of thymosin fraction 5 and thymosin α1. J Biol Response Mod 1:35-41, 1982.
9. MALUISH AE, ORTALDO JR, SHERWIN SA, ET AL. Changes in immune function in patients receiving natural leukocyte interferon. J Biol Response Mod 2:418-427, 1983.
10. HENGST JCD, KEMPF RA, KAN-MITCHELL J, ET AL. Immunological effects of recombinant interferon-α2 in cancer patients. J Biol Response Mod 2:516-527, 1983.

GENERAL DISCUSSION: Biological Response Modifier Trials

Q: (Dr. Richard Simon, NCI) I would like to make three comments on these presentations. First, dose-response experiments where the maximum effect occurs at the lowest dose are of concern to me because you may not be seeing any effect, except for a negative effect at the higher doses. I think one needs to consider having a zero dose control. Second, I was wondering about what kind of dose-response relationships are seen in animal systems with biological response modifiers (BRMs); do you find that intermediate doses give the most effect, or the lowest doses, or the highest doses? Finally, two of the designs that were discussed yesterday—the response surface designs to find optimal combinations of doses and factorial designs, where you superimpose an additional randomized treatment assignment on an existing one without gross increases in the number of patients involved—might find very good applications in the development of BRMs.

A: (Dr. Ronald Herberman) In animal tumor studies where one has much greater flexibility of evaluating a wide range of doses, the experience with several biological response modifiers has been that a bell-shaped dose-response curve is found. I think your points about the clinical studies are well-taken. I wouldn't be satisfied to settle with a result in which the best results were at the lowest dose, just as I wouldn't be satisfied with that in an in vitro experiment. I'm not sure whether that would necessitate going to zero but it would certainly, I think, necessitate going down substantially lower to see if you can get below those biologic effects.

A: (Dr. Annette Maluish) I was just going to make a similar comment about the very low doses. In the Poly IC-LC study, there's a clear increase at quite a low dose, but I would agree with Dr. Herberman that we would take it, not necessarily to a zero dose, but to very, very low doses, and, in fact, that's the aim of what we're trying to do now with the alpha-interferon studies. I think we also have to address the question of the frequency and the timing of administration in those studies.

A: (Dr. Carl Pinsky) I think it's important to point out that it's not just biologic effects, but also antitumor effects which demonstrate the same sort of bell-shaped dose-response curve in the animal model systems. If we want to test the hypotheses that we set out to, we must try to work at the optimal BRM dose.

Q: (Dr. Brigid Leventhal, Johns Hopkins) I wanted to essentially support what Dick Simon said. I think this is a situation where you almost have to have a 2×2 factorial design because you really need an untreated control, and you could legitimately do that. You need it for Dr. Maluish's studies where the effect of immune response of growing tumor needs to be controlled for. Because it's possible to enhance the growth of tumors with immunotherapy, I think that you should have an untreated control. And the final point, as we all know, is that with combinations you may abrogate the immune effect with the cytotoxic agent and so you definitely need the separate elements as treatment groups, you need it even more. You need these controls in chemotherapy, too, but the enhancement question is one we have to come to grips with particularly in immunotherapy, and that means an untreated control.

A: (Dr. Herberman) I think those points are well-taken, Brigid.

Q: (Dr. Robert Dillman, San Diego) It seems that one of the biggest problems we have had is that we have to start off with an initial phase I design to answer toxicity questions, and we have to use numbers of patients appropriate for such a trial. And yet, our goal is more a classical phase II efficacy question, in terms of the numbers of patients needed to determine biological response modification. By the nature of the very complex endpoints we are dealing with, using assays that are very difficult to standardize, we have been forced to rely on single-institution studies. We run into the problem of patient numbers

that would really be needed to get any kind of a statistically valid result in terms of biological response modification. As Carl pointed out, in the study we performed, it was actually a phase I design from which we tried to draw phase II conclusions with insufficient numbers. From the standpoint of the Biological Response Modifiers Program and the NCI, I wonder what approaches may be to get around this.

A: (Dr. Herberman) I think that's a good point. We don't have a perfect answer at the moment; we're wrestling with it right now. I think, at least with certain tumor types, a proper design could be achieved now without an inordinate expense or number of patients. At least I think we can do considerably better right now with the available resources than we've done until now.

A: (Dr. Pinsky) During phase Ia, as you go through the dose escalation scheme, you get certain hints about where you may be getting activity. As has been pointed out, the numbers are extremely small if you treat only three or four patients at each dose level. I think what you need to do at the tail end of any one of those phase I trials is to start to really direct yourself toward the phase II design, that is, to treat additional patients, probably no less than 15 or 20 patients at any given dose level, to really get consistent and statistically significant changes in these assays that we're measuring. And then I think you would be a lot more confident about going to phase II with the optimum BRM dose.

Q: (Dr. Herberman) I would like to raise a question about the hypotheses that Dr. Pinsky started his talk with. Essentially, the first hypothesis or two that he listed were a version of the immune surveillance hypothesis, ie, that deficits in some part of the immune response might lead to the development or progression of tumors, and if you corrected that, you'd expect this would lead to regression of the tumor. But I think, in fact, for some of the effector mechanisms that the biological response therapy is based upon, there might not be deficits in the usual sense of being depressed below normal. I would personally believe that there would be the potential that, although some of these responses are normal, the same as someone without a tumor, they might still be insufficient to deal with the tumor. There would still be an alternative hypothesis, that if you made some of these reactivities hypernormal, you might then get regression of tumor. Perhaps you could comment on that modification of your hypothesis, Carl.

A: (Dr. Pinsky) I think it has been well shown that not all assays of immunocompetence have been measurably abnormal in patients with cancer, sometimes even with advanced cancer. It's remarkable how well some responses are maintained even when the cancer becomes far advanced. I think my admittedly simplified first slide was intended to say that even the type of relative deficiency that you described would be a defect that you could attempt to correct with a BRM. And it wouldn't necessarily be an abnormality that would simply be compared to normals, but rather something that would be clearly related to disease progression and perhaps metastasis, and, as such, exploitable or attacked by a BRM.

(Dr. Pinsky) I wanted to close with a comment to any of the biostatisticians who have looked at some of this and might be as impressed as I am with the amount of gold there is to mine in this particular area. The questions of dose-response, multiple assays, changes with time, whether or not untreated control groups are necessary, and so forth, are extremely important questions that we have to ask. And many of these questions have only been touched as far as scratching the surface—most of these questions have not been really analyzed in detail. This sort of help is desperately needed.

(Dr. Herberman) I would completely second that. I was really surprised when we got into some of these issues with Dr. Maluish. We went to the literature or talked to statisticians and expected to find very well-developed methods to utilize with these systems, to find that they just didn't exist, and they needed to be developed rather painstakingly.

Site Visit Monitoring Program of the Clinical Cooperative Groups: Results of the First 3 Years

Joan K. Mauer,[1] Daniel F. Hoth,[1,*] Dorothy K. Macfarlane,[1] Luz D. Hammershaimb,[2] and Robert E. Wittes[1]

As the world's largest sponsor of clinical trials of investigational antineoplastic agents, the National Cancer Institute (NCI) must ensure that research data generated under its sponsorship are of high quality, reliable, verifiable, and reproducible. NCI's monitoring policies for clinical research have been in evolution since the start of cooperative cancer trials in the late 1950s. One important instrument of quality control has always been that investigators in the cooperative group and contract programs have undergone peer review as part of the funding process. As the NCI's clinical research program increased in size and complexity, the systems for quality control became more formal and systematic. In 1963, Congress passed the Harris-Kefauver amendments to the Food, Drug and Cosmetic Act requiring the Food and Drug Administration (FDA) to oversee Investigational New Drug (IND) testing in human subjects. In 1977, the FDA published proposed regulations on the responsibilities of sponsors and monitors of clinical trials. Although not published in final form until 1985, this proposal, calling for an annual site visit to each investigator, had a profound effect on the sponsors of investigational drug trials in the United States. Most sponsors changed their practices to conform to these proposals.

In 1978, the Eastern Cooperative Oncology Group began routine site visits to verify accuracy of submitted data for its clinical trials. In 1979, the NCI established the Clinical Trials Monitoring Service as a formal data base management and site visit program for phase I trials. In 1982, the NCI made site visit monitoring a requirement for all cooperative groups, cancer centers, and any other investigators conducting clinical trials under its IND sponsorship.

In this paper we describe in detail the monitoring programs of the clinical cooperative groups and present the results of the first 3 years of site visits.

BACKGROUND

The Clinical Cooperative Group Program is a major component of the extramural research effort of the NCI's Division of Cancer Treatment. Each of the 24 individual clinical trials groups represents a consortium of institutions working through an elected and appointed leadership. Some groups restrict their attention to simple disease sites, while others approach many types of neoplasms with a broad array of therapeutic maneuvers.

The groups are organized around a headquarters, led by an elected group chairman, and a statistical office. The voting membership is usually comprised of principal investigators at full member institutions. Many groups also include affiliate members, which are hospitals, clinics, or individual physicians closely associated with a member institution. Affiliation provides the group with increased patient accrual and also involves community physicians in clinical research. Each group has developed policies and criteria for both members and affiliates.

Since the multicenter nature of group trials presents obvious questions about variability, the cooperative groups long ago recognized the need for formal quality control and monitoring. Procedures were developed for monitoring the overall progress of studies and for ensuring adherence to protocol and procedural requirements. More recently, the importance of assuring the accuracy of data by comparing submitted data to the primary medical record has led to formal site visit monitoring within the cooperative groups. Because quality control and assurance programs already were in place in many groups, the NCI delegated its responsibility for monitoring investigational drug studies to the groups, as described below.

The groups perform two distinct kinds of monitoring. The first is periodic review of the overall progress of each

[1]Cancer Therapy Evaluation Program Division of Cancer Treatment, National Cancer Institute, Bethesda, MD.
[2]Lederele Laboratories, Pearl River, NY.
*Reprint requests to: Dr. Daniel F. Hoth, Investigational Drug Branch, Cancer Therapy Evaluation Program, Division of Cancer Treatment, National Cancer Institute, Landow Bldg, Rm 4C09, National Institutes of Health, Bethesda, MD 20205.

study to assure that the projected accrual goals are met on a timely basis, that overaccrual is avoided, that eligibility and evaluability rates do not fall below minimum acceptable standards, and that risks do not outweigh benefits. The cooperative groups perform this function at least semiannually prior to their group meetings.

Also included in these central quality control measures is the assessment of protocol compliance. This is done in an increasingly systematic way and on a real-time basis. For example, most cooperative groups conduct central pathology review for selected studies to reduce variability in diagnosis. To ensure adherence to protocol-specified treatment, radiotherapy films and surgery reports are also monitored centrally. Checks of submitted flowsheets for protocol compliance ensure that drug treatment is delivered according to protocol stipulations and that appropriate study tests have been obtained. The study chairman and/or the statistical office are responsible for determining each case's eligibility and evaluability, based on the information gathered through these quality control mechanisms.

The second type of monitoring assures the quality of trial execution at the level of the investigator. The audit program assures that the data used to analyze the trials are an accurate reflection of the raw data. The program requires an on-site comparison of the submitted data with the primary medical record for a sample of patient cases. At the same time, compliance with regulatory requirements for the protection of human subjects and investigational drug accountability are checked.

The NCI maintains oversight responsibility for the group's monitoring program. The NCI has provided guidelines for the establishment of monitoring programs to each group and help in the implementation of each group's audit program. NCI staff reviews all site visit schedules and all reports of audit findings. To assure consistency of monitoring across the groups, an NCI representative attends 10%–20% of all cooperative group site visits. Quality assurance staff may make specific recommendations for action if they do not feel the action taken by the group has been adequate.

DESCRIPTION OF THE PROGRAM

Scheduling of Audits

All cooperative group member and affiliate institutions are audited at least once every 3 years. Institutions are scheduled for audit randomly by the group but each is at yearly risk of being audited. The lead time for informing the institution of selected patient cases is kept to a minimum.

Audit Preparations

The cooperative group's audit preparations include the following steps:

1. The statistical office of the cooperative group selects at least three protocols for review. These include, if applicable: (*a*) IND protocols, (*b*) multimodality protocols, (*c*) pilot studies, and (*d*) intergroup studies.
2. The statistical office randomly chooses four to eight cases, or approximately 10% of entered patients per protocol, for review. A total of 20–30 cases are chosen.
3. The site visit team audits the flowsheets submitted by the institution to the statistical office.
4. The site visit team generally consists of two group investigators and two data managers. The team may be augmented by statistical office representatives and others, including NCI observers.
5. Approximately 2 weeks in advance of the visit, the institution is supplied with the list of selected patient cases.

The institution prepares for the audit by gathering all source documentation pertaining to the selected cases. For each selected case the following records should be available: informed consent document; protocol flowsheets; hospital charts; physician and research notes; outpatient and clinic records; and x-rays, scans, and other pertinent studies.

For the selected protocols, the data should include: (*a*) the Institutional Review Board (IRB) approvals and reapprovals; (*b*) annual reports submitted to the IRB; and (*c*) the current version of the protocols in use at the institution and any amendments. Finally, all records regarding the disposition of investigational drugs, specifically the NCI Drug Accountability Records, should be available.

Conduct of Audit

The auditors review specific data relating to research and regulatory requirements, as shown in detail in the appendix. At the conclusion of the visit, the audit team conducts an exit interview with the investigator and his staff to review the findings of the audit and to discuss recommendations. This interview provides the opportunity for immediate dialogue and feedback and/or clarification by personnel at the audited institution.

A written report of the audit findings is submitted to the group audit committee and leadership for review and action, if required. Many of the groups have integrated the audit findings with the central quality control evaluations described earlier to establish a final decision on eligibility and evaluability for each case reviewed. Most of the groups also incorporate the audit results into their periodic performance review of each institution by the membership committee.

In audits where the findings indicate poor data quality or noncompliance with regulatory requirements, the group may take a variety of actions depending on the

scope and severity of the problem:

1. The institution's staff is notified of the problems encountered during the audit and instructed in ways to improve performance.
2. A reaudit may be performed within a time frame shorter than the usual 3-year cycle to evaluate improvement.
3. In cases where severe or multiple problems are noted, patient registration privileges may be suspended until a larger sampling of cases can be audited or until the group is assured that internal procedures at the institution have been modified to prevent recurrence of these problems.
4. If the group is not satisfied that the problems are correctable, it may choose to terminate the membership or affiliate status of the institution.

MATERIALS AND METHODS

We present an analysis of data collected over a 3-year period from January 1982 through December 1984 in the 17 groups conducting clinical trials with experimental agents. The results of audits of 812 institutions, 2814 protocols, and 5988 patient cases form the basis of this report.

We have estimated the quality of performance in the several relevant categories as the fraction of audited elements which have been judged deficient. The numerator represents those deficiencies which meet the criteria for each audit element as described below. The denominator is the total of all group audit findings in each category; accordingly, the denominator can be patient cases, institutions, or protocols. Performance in the area of drug accountability is calculated by institution.

If $\geq 30\%$ deficiencies are noted in patient cases or protocols at an institution, that institution is tabulated as deficient for the institutional performance analysis.

Many submitted reports of audit findings from the cooperative groups did not address certain key elements of an audit. Only those elements (ie, eligibility, treatment, etc) clearly stated in the report of the audit team were used for this analysis. When an element was not included in the report, that element was subtracted from the total denominator (ie, protocols, patients, or institutions). If patient cases or protocols reviewed at an institution met the criteria described below, they were included in the analysis. The data derived from audit reports were analyzed separately for member and for affiliate institutions. For reports that did not clearly identify affiliate institutions, number of protocols, or affiliate cases reviewed, these results were included in the member institution's data. The results are tabulated by the following criteria.

Research Elements

Eligibility.—Only patients judged clearly ineligible by the auditors are considered deficient in this category. Cases where eligibility was doubtful or not mentioned are not included.

Treatment inconsistencies.—Deficiencies include $\geq 10\%$ deviations in doses administered, incorrect drug, incorrect schedules, or doses or schedule modifications not justified by the protocol or concomitant therapy known to affect protocol treatment. Wherever a site visit report alluded to treatment which differed from that stipulated by protocol, the institution was judged deficient in this category.

Response assessment.—Response assessment was considered deficient if the judgment of the auditor differed from that of the investigator. Problems with response assessment without a patient breakdown were also included. Examples of undocumented response include lack of tumor measurements at critical time points.

Toxicity assessment.—Deficiencies for toxicity assessment and reporting consist of (*a*) unreported life-threatening adverse drug reactions such as renal or hepatic failure; (*b*) unreported significant myelosuppression leading to sepsis; (*c*) an overall impression of consistant undergrading of toxic effects; and (*d*) a general description of deficiencies in this area as stated in the audit report but without specific numbers.

Data verification.—This category was judged deficient in cases for which the auditors were unable to verify the existence of data on critical study parameters (ie, those which assure end points, including blood cell counts, or follow-up data when a patient is discharged), or for which there was frequent disagreement between submitted data and the medical record. Indications of inconsistent performance in this area as noted in the audit report without specific numbers were also included in the deficiencies category.

Regulatory Elements

IRB.—Deficiencies scored in this analysis included: (*a*) missing IRB approval documentation for one or more protocols, and (*b*) IRB approval after patient entry for one or more protocols. If the report was not specific but alluded to problems with IRB approvals, these were also considered deficiencies.

Informed consent.—Patient cases were included in the deficiency category if one or more of the following were found: (*a*) missing documentation of one or more informed consent documents; (*b*) substantial delays (> 48 hours) in the patient signing consent form after initial treatment; and (*c*) use of consent forms not specific to the protocol. If the report cited significant problems with informed consent documents but did not include actual numbers, these were also included.

Drug accountability.—The criterion for inclusion in deficiencies was that one or more of the following were found: (*a*) NCI Drug Accountability Records were not in

use; (*b*) NCI Drug Accountability Records were not protocol-specific; (*c*) separate drug stocks were not maintained for each protocol; (*d*) drug storage or security departed from NCI guidelines; (*e*) there was drug leakage to unapproved protocols; or (*f*) there was an inventory discrepancy. The requirements for the use of NCI Drug Accountability Records went into effect February 1983. Therefore, violations are included only from June 1983 to December 1984.

Cooperative Group Response to Site Visit Findings

The actions taken by the cooperative group in response to reports of its site visit teams reflect its judgment concerning the seriousness of the problems encountered. Where several actions ensued, only the most serious was tabulated. In other words, institutions suspended from entry privileges which were also issued warning letters were considered in the "under suspension" category only.

RESULTS

Audit reports from January 1982 to December 1984 reveal that 17 cooperative groups were actively site-auditing. As of December 1984, 812 institution participants in cooperative group trials have been visited. An institution may be visited several times because individual units in the institution may participate in more than one cooperative group. The total of 812 institutions encompasses both main member and affiliate member audits. An overall summary of the results is shown in table 1.

Table 1

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM
SUMMARY OF RESULTS
1/82 - 12/84

# of Groups	17
# of Institutions Audited	812
# of Protocols Sampled	2814
# of Cases Reviewed	5988

	Cases Eval		Deficiencies By Case		Institutions Eval		Deficiencies By Institution	
	N	%	N	%	N	%	N	%
Research Elements								
Eligibility	4814	80	346	7	652	80	46	7
Treatment Deviations	4854	82	595	12	662	82	126	19
Response Assessments	4553	76	284	6	555	68	58	10
Toxicity Reporting	4786	80	169	4	529	65	47	9
Data Verification	4790	80	260	5	662	81	87	13
Regulatory Elements								
IRB*	2692	95	168	6	670	83	79	12
Informed Consent	5023	84	455	9	696	86	83	12
Drug Accountability					567	70	79	14

* Per protocol

Case Eval. N = # of patients evaluable for analysis from the submitted report.
Deficiencies by Case N = # of cases considered by criteria to be deficient.
Institutions Eval. N = # of institutions evaluable for analysis from the submitted report.
Deficiencies by Institutions N = # of institutions considered by criteria to be deficient.

REGULATORY COMPLIANCE

Institutional Review Board

Table 2 shows that IRB problems as described above were found for 168 (6%) protocols reviewed.

Informed Consent

Four hundred fifty-five cases from all institutions representing 9% of those with evaluable data showed deficiencies with informed consent requirements. Institutions where ⩾ 30% of the cases reviewed did not meet these requirements constituted 12% of those audited (table 3).

PATIENT CASE AUDIT RESULTS

Eligibility

Of 5988 patients reviewed, there were reports about the assessment of eligibility in 80% of patient cases. Seven percent were clearly ineligible. Seven percent or 46 institutions were found to have a ⩾ 30% ineligibility rate. Whether these cases had been previously excluded through the central quality control mechanisms of the group was not recorded in most audit reports (table 1).

Treatment Variations

Of the 4854 cases for which treatment variations could be assessed, 12% were noted to have treatment deficiencies significant enough to render them unevaluable for protocol analysis. In 126 (19%) of the 662 member institutions with adequate reports to evaluate, ⩾ 30% of the cases demonstrated significant departures from protocol-stipulated treatment (table 1).

Response

Lack of response documentation or disagreement with response assessment in ⩾ 30% of cases was found in 58 of 555 evaluable member institutions, while 6% of 4553 evaluable cases overall demonstrated deficiencies of this nature (table 1).

Toxicity

Forty-seven or 9% of member institutions showed significant deficiencies in toxicity reporting. Overall case analysis revealed that 4% of the 4786 evaluable cases had problems in this element (table 1).

Data Verification

Eighty-seven institutions (13%) were rated as having

Table 2

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM RESULTS
All Institutions - 1/82 - 12/84
Deficiencies By Case Analysis

	1982		1983		1984		TOTAL	
	N	%	N	%	N	%	N	%
Research Elements								
Eligibility	86/994	9	149/2188	7	111/1632	7	346/4814	7
Treatment	186/949	20	227/2456	9	182/1549	11	595/4954	12
Response Assessments	87/923	9	96/1838	5	101/1792	6	284/4553	6
Toxicity Reporting	43/1063	4	61/2137	3	65/1586	4	169/4786	4
Data Verification	127/1124	11	84/2231	4	49/1435	3	260/4790	5
Regulatory Elements								
IRB*	46/307	15	97/1361	7	25/1024	2	168/2692	6
Informed Consent	91/1082	8	200/2165	9	164/1776	9	455/5023	9

* Per Protocol

N = # of patient cases (IRB/protocol) reviewed that meet the deficiency criteria for each element.

Table 3

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM RESULTS
Deficiencies By All Institutions - 1/82 - 12/84

	1982	1983	1984
# of Groups	10	14	17
# of Institutions	135	404	273
# of Protocols	414	1367	1033
# of Cases	1297	2816	1875

	1982		1983		1984		TOTAL	
	N	%	N	%	N	%	N	%
Research Elements								
Eligibility	12/15	10	19/317	5	15/220	7	46/652	7
Treatment Deviations	27/115	23	54/317	17	45/230	20	126/662	19
Response Assessments	20/106	19	16/255	6	22/194	11	58/555	10
Toxicity Reporting	12/102	12	15/278	5	20/149	13	47/529	9
Data Verification	19/114	16	48/327	15	20/221	9	87/662	13
Regulatory Elements								
IRB*	15/77	20	47/284	17	17/309	6	79/670	12
Informed Consent	12/107	11	39/356	11	32/233	14	83/696	12
Drug Accountability			30/307	10	48/176	27	78/567	14

N = # of institutions meeting the criteria for ≥30% deficiencies per audit element.

≥ 30% of the cases lacking documentation of critical data in the primary medical record or showing discrepancies between the medical record and the research record. Five percent of the cases reviewed revealed problems of this nature (table 1).

Drug Accountability

Since drug accountability audits by the cooperative groups were not performed during the entire time period of this review, the number of reports dealing with drug accountability is low. If only adequate reports are considered (567 evaluable institutions), 14% had drug accountability problems. Since the audits of affiliate institutions are not always "on site," drug accountability was not always reviewed (table 1).

Responses by Groups to Site Visit Findings

For main member institutions, 54 (13%) were revisited earlier than the usual 3-year cycle based on the findings of the original audit; 52 affiliates or 13% were revisited (table 4). In some instances, institutions were placed on probation: nine (3%) members and 31 (8%) affiliates. Two main members were dropped and 11 (3%) affiliate institutions were dropped because of audit findings.

Affiliate Analysis

In 1982, during routine audits of affiliate institutions, two separate cooperative groups found that two affiliates had submitted falsified data. The total of 56 patient cases was misrepresented. The groups immediately dropped the affiliates and expunged all data submitted by them. The NCI suspended their access to investigational agents. The FDA conducted independent audits of both affiliates and, among other actions, revoked their investigator privileges for investigational agents. In light of these disturbing findings, the NCI asked in early 1983 that all groups accelerate the auditing of all their affiliates by completing the initial audit of each affiliate within the next 18 months.

Table 4

COOPERATIVE GROUP SITE VISITS MONITORING PROGRAM
Actions for Problems - 1/82 - 12/84

Type of Actions	1982 N	1982 %	1983 N	1983 %	1984 N	1984 %	TOTAL N	TOTAL %
MAIN MEMBERS								
Letters	9/84	11	10/184	5	20/152	13	39/420	9
Re-Site Visit	6/84	7	24/184	13	24/152	16	54/420	13
Suspend/Probation			6/184	3	3/152	2	9/336	3
Dropped			2/184	1	0/152	0	2/336	1
AFFILIATE INSTITUTIONS								
Letters	11/50	22	5/220	2	5/121	4	21/391	5
Re-Site Visit	10/50	20	20/220	9	22/121	18	52/391	13
Suspend/Probation	4/50	8	26/220	12	1/121	8	31/391	8
Dropped	3/50	6	8/220	4	0/121	0	11/391	3
Others, e.g., FDA Action	2/50	4	0/220	0	0/121	0	2/391	1

N = # of institutions that groups took actions against.

Table 5 illustrates the groups' efforts in 1983 (220 affiliates audited) and in 1984 (121 affiliates audited). Overall, the performance of affiliates has improved dramatically. The data show that ineligible cases declined from 16% (1982) to 8% (1984), treatment deficiencies declined from 62% to 19%, and unsupported response assessments decreased from 17% (1982) to 10% (1984) (table 6). Analysis of the regulatory elements of the audits shows that in 1982, 45% of institutions had problems with IRB approval requirements. By 1984, the proportion of institutions with such deficiencies decreased to 6% (table 5). Departures from informed consent requirements by case decreased from 26% (1982) to 12% (1984) (table 6).

Table 4 illustrates that action was taken by the groups regarding affiliate performance in approximately 30% of the institutions audited. Five percent of the affiliates received warning letters, 13% were reaudited earlier than the usual 3-year cycle, 8% were placed on probation or suspended, and 3% were dropped from group affiliation.

Tables 7 and 8 are comparisons by year of member and affiliate institutions. Table 7 clearly shows that in 1982 affiliate institutions had far more deficiencies in both regulatory requirements and research elements than did member institutions. In the two succeeding years, the gap in performance has narrowed considerably.

Table 8 compares the performance of members and affiliates by cases reviewed. Again, the very large differences found in 1982 for both research and regulatory elements have become less impressive with the passage of time.

DISCUSSION

Three years after its inception, the monitoring program for the cooperative groups has accomplished site visits to ⅔ of its main member institutions and ½ of affiliate institutions. Approximately 10% of the total patient accrual from these group members has been subjected to an unprecedented in-depth scrutiny for regulatory compliance, protocol adherence, and assessment of data quality.

All groups did not begin monitoring in 1982, since some were not then being supported by cooperative agreements. The 3-year member cycle and 18-month affiliate cycle are different for each group. Some groups completed their cycles in 1984, while others will not complete theirs until 1987.

For each element of audit findings, we set the threshold rate of deficiencies for inclusion in the tabulation at ≥ 30%. This figure was intentionally set at a high level to highlight serious instances of deficiencies in the execution of clinical trials.

Analysis of member institutions shows a generally satisfactory level of performance, with a small and constant level of poor performers over the 3-year period. By

Table 5

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM
AFFILIATES - 1/82 - 12/84
Deficiencies By Institutional Performance

	1982	1983	1984
# of Institutions	50	220	121
# of Protocols	79	236	232
# of Cases	228	670	335

	1982			1983			1984		
	N	%	Problem %	N	%	Problem %	N	%	Problem %
Research Elements									
Eligibility	9/43	21		10/164	6		6/85	7	
Treatment Deviations	15/42	36		40/168	24	1 5	25/97	26	
Response Assessments	11/35	31		9/126	7		14/70	20	
Toxicity Assessment	8/33	24		9/122	7	2 6	9/76	12	
Data Verification	10/38	26	1/38 3	24/158	15	7/158 4	6/81	7	
Regulatory Elements									
IRB			11/24 45			36/145 25			11/119 6
Informed Consent	10/36	28		28/181	15		20/87	23	
Drug Accountability						13/140 9			22/40 55

N = # of institutions meeting the criteria of ≥30% deficiencies.

"Problem" = the audit report did not give specifics but alluded to deficiencies.

contrast, a sharp improvement has been seen in the performance of affiliate institutions. This may have been the result of either increased educational efforts by the groups or the voluntary dropping of many affiliates which were small accruers. The groups have drawn up formal policies for the acceptance of affiliates and for the performance expected from these institutions to retain affiliate status. In any case, the data from affiliate institutions are clearly more reliable and accurate in 1984 than they were in 1982.

Certain findings from the monitoring program have already had a general salutory impact on clinical trials practices. For example, with respect to variability in response assessment, the monitoring program of one cooperative group has found that for certain diseases such as lung cancer, assessment of response by audited investigators varied significantly from that of the audit teams. Variability in response assessment is, of course, not a new finding (1,2). The site visit monitoring program has increased general awareness of this potential variable and in fact provides an excellent opportunity for estimating its prevalence across diseases. The finding of substantial interobserver variability in many studies would have important implications for the utility of response as a primary endpoint in cancer trials.

Some audit programs have revealed that increasing complexity of study design leads to lower compliance rates. These findings have provided the impetus for simplification of protocol design in subsequent studies.

The site visit program has also enabled the groups to gauge the completeness of submitted data. Some groups have utilized site visits to retrieve missing data on the flowsheet from source medical records. Obviously, the more complete the information, the more valid a study's conclusions are likely to be.

The monitoring program has had an important educational function which has been widely acknowledged by many groups. Investigators have become more aware of the need to maintain well-documented research records. Site visitors and investigators have also become more familiar with regulatory requirements. In most of the groups a large segment of investigators and data managers participate as auditors. After auditing other institutions, participants recognize ways to improve their own research practices.

Experience to date, therefore, suggests that the moni-

Table 6

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM
AFFILIATES - 1/82 - 12/84
Deficiencies By Case Analysis

	1982		1983		1984	
	N	%	N	%	N	%
Research Elements						
Eligibility	34/210	16	39/474	8	24/314	8
Treatment Deviation	102/165	62	90/568	15	59/307	19
Response Assessments	32/186	17	5/403	1	28/294	10
Toxicity Reporting	14/168	8	17/514	3	21/305	7
Data Verification	97/206	47	46/537	9	10/283	4
Regulatory Elements						
IRB*	35/50	70	72/320	22	14/225	6
Informed Consent	47/180	26	52/589	8	37/304	12

* Per Protocol

N = # of cases meeting the criteria for deficiencies

Table 7

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM
COMPARING MEMBERS AND AFFILIATES BY INSTITUTIONAL PERFORMANCE
1/82 - 12/84

	1982				1983				1984			
	Mem.		Aff.		Mem.		Aff.		Mem.		Aff.	
	N	%	N	%	N	%	N	%	N	%	N	%
Research Elements												
Eligibility	2/72	3	9/43	21	9/153	6	10/164	6	7/135	5	6/85	7
Treatment Deviations	10/73	14	15/42	36	13/149	9	40/168	24	17/133	13	25/97	26
Response Assessment	6/71	8	11/35	31	7/129	5	9/126	7	2/124	6	14/70	20
Toxicity Assessment	2/69	3	8/33	24	4/146	3	9/122	7	9/73	12	9/76	12
Data Verification	2/76	3	10/38	26	4/169	2	24/158	15	6/140	4	6/81	7
Regulatory Elements												
IRB	4/53	8	11/24	45	11/139	8	36/145	25	7/140	5	11/119	6
Informed Consent	2/71	3	10/36	28	11/175	6	28/181	15	12/146	8	20/87	23
Drug Accountability					17/167	10			26/136	19		8

N = # of institutions meeting the analysis criteria

Table 8

COOPERATIVE GROUPS SITE VISIT MONITORING PROGRAM
COMPARING MEMBERS & AFFILIATES BY CASE
1/82 - 12/84

	1982				1983				1984			
	Mem.		Aff.		Mem.		Aff.		Mem.		Aff.	
	N	%	N	%	N	%	N	%	N	%	N	%
Research Elements												
Eligibility	52/784	6	34/210	16	110/1714	6	39/474	8	87/1318	6	24/314	8
Treatment Deviations	84/784	11	102/165	62	137/1524	9	90/568	15	123/1242	10	59/307	19
Response Assessment	55/737	7	32/186	17	91/1435	6	5/403	1	73/1204	6	28/294	10
Toxicity Assessment	29/727	4	14/168	8	44/1623	3	17/514	3	44/1281	3	21/305	7
Data Verification	30/918	3	97/206	47	38/1694	2	46/537	9	39/1152	3	10/283	4
Regulatory Elements												
IRB*	11/257	4	35/50	70	25/1041	2	72/320	22	11/797	1	14/225	6
Informed Consent	44/902	5	37/180	26	148/1576	9	52/589	8	127/1472	9	37/304	12

* Per protocol

N = # of patients cases meeting the criteria for deficiencies

toring program can augment quality control measures performed centrally, since on-site audits provide access to information found only in the primary medical record. The program has also been successful in identifying problems with the informed consent process and IRB approvals, procedures which are not usually monitored centrally. The monitoring program also provides the opportunity for rapid feedback to the investigators and data managers to correct misconceptions and problems uncovered during the site visit. As a result of audit findings, investigators have recognized the need to establish clear policies and standards for the acceptance of affiliate institutions and requirements for maintenance of affiliate status.

Based on the data reviewed here, we believe that the cooperative clinical trials network is fundamentally healthy and conducts research according to high standards. Most of the problems appear correctable through education of investigators and data managers. However, the data do not give grounds for complacency, inasmuch as approximately 17% of member institutions and 25% of the affiliate institutions warranted some corrective measures in the judgment of the groups. We should also acknowledge that the use of 30% as the cutoff figure for categorizing an institution as deficient may yield an impression that all is well in certain categories, when in fact it may not be. Thirty percent or more deficiencies is an extremely high figure for any category; the fact that an institution may fall below this level (and therefore is not considered deficient in our analysis) is not necessarily cause for rejoicing.

In short, the data indicate that continued adherence to the details of clinical trials performance and data quality will be necessary in the future. The site visit monitoring program will provide the means for tracing performance for individual institutions and groups and will be a useful tool for the assessment of time trends in performance. In addition, although the monitoring program had its origins in regulatory necessity, the activity is far more than a surveillance activity. Data gathered in properly prepared and performed audits are a unique window on certain aspects of clinical trials methodology. We hope that the cooperative group members who devote much of their time to this effort will recognize that this activity can tell us much about the clinical research activity itself. In this respect, therefore, the results of the monitoring program are as worthy of thought and attention as the trials themselves.

APPENDIX

1. Research Elements Review
   Comparison of submitted data to primary medical records to verify:
   a. Patient eligibility
   b. Protocol treatment administered as per the submitted flowsheets
   c. Tumor response and its assessment via review of x-rays, scans, etc, and physical examinations
   d. Reported toxicity and its assessment per protocol
   e. Required laboratory and diagnostic evaluations are actually performed
   f. Overall quality of record keeping

   Also to observe the records for patient information not included in submitted research record, eg, concomitant therapy
2. Regulatory Requirements Review
   a. Documentation of IRB approvals, reapprovals, and annual IRB reports for each audited protocol and its amendments
   b. Documentation of an IRB-approved, properly signed and dated informed consent document for each case audited
   c. Inspection of the adequacy and security of investigational drug handling
   d. Documentation of proper use of NCI Drug Accountability Records
   e. Comparison of shelf inventories for several drugs versus the balance of drug on the NCI Drug Accountability Records
   f. Comparison of specific NCI Drug Accountability Record entries to the patient source documentation to cross-check actual doses dispensed and dates of administration

REFERENCES

1. WARR D, MCKINNEY S, and TANNOCK I. Influence of measurement error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. J Clin Oncol 2:1040–1046, 1984.
2. MOERTEL CG, and HANLEY JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. Cancer 38:388–395, 1976.

Quality Assurance in a Cooperative Group

Janet M. Wolter[1,*]

Quality control in a large cooperative group involves a number of different levels of checking which can be divided into three main sections: quality control in data management, quality control in special areas, and the verification that events on paper reflect real events in the clinic. Quality control in special areas is the province of those using various techniques and disciplines and it will not be discussed in this overview, since others will cover the monitoring of radiation therapy, standardization of biochemical tests, and pathology review. [Cancer Treat Rep 69:1189–1193, 1985]

Quality Control in Data Management

Quality control in data management begins at the institutional level, with the ultimate validity of the data depending on the adequacy of institutional data handling. In the Eastern Cooperative Oncology Group (ECOG), each data manager is specially trained, there are numerous workshops held for data managers at all levels, and an extensive Data Managers' Manual covers almost every imaginable situation. In the ECOG, main institutions are responsible not only for data on their own patients but also for data on patients accrued from satellites as well. Flow sheets are reviewed page by page before being submitted to the Statistical Center. Original forms are always sent, and any special material such as slides, radiation therapy port films, etc, are all sent to the Operations Office where they are logged in and forwarded to the appropriate area for recording. Figure 1 is a schematic of how these materials are handled and how institutions are involved in follow-up. Before any material is sent, protocol items reviewed by data managers in the institutions include eligibility; checks on treatment administration; a review of the records to be kept, especially specific forms for specific diseases; a table of parameters to be followed; and the presence of any new forms being tested by the cooperative group. It is particularly important that investigators recognize that excellence of completed studies is dependent on the training and efforts of local data managers. In the ECOG there is an active Data Manager Committee, members of which participate in forms development and are active at the meetings of Disease-Oriented Committees. In addition to workshops for data managers, frequent educational programs targeted at various levels are presented to make sure that the data managers have as much knowledge as possible about the diseases and the treatments with which they are dealing. In addition, the Chief Data Manager frequently "circuit rides" to help on the spot with any problems that might have arisen. Funds are set aside to permit new data managers to travel to nearby institutions with good data management should there be no immediate workshop available.

While quality begins at the institutional level, the Statistical Center data managers are responsible for specific studies. They check the forms they receive for consistency and generate, if necessary, query letters which are returned to the institutions. These query letters are on a single sheet of paper with contact carbon so that the question and the answer can be filed by both the Statistical Center and the institution. The number of query letters necessary to clarify submitted data for any given study or time interval is a crude measure of institutional efficiency. The Statistical Center data manager then does a case evaluation and forwards the material to the study chairman, who has the final say on eligibility/evaluability. Both the data manager and the computer check compliance on drug doses, monitor blood counts, and evaluate toxicity as reports flow in. All the data are reviewed both by the data manager and the study chairman as each step of a protocol is completed. Whereas compliance has been heavily emphasized as something to be checked in on-site audits, the ECOG computer check immediately calls attention to failure of compliance as the printouts are generated. Figures 2 and 3 are samples of a query letter and a toxicity report.

[1]Section of Medical Oncology, Rush-Presbyterian-St. Luke's Medical Center, Chicago, IL.

Reprint requests to: Janet M. Wolter, MD, Section of Medical Oncology, Rush-Presbyterian-St. Luke's Medical Center, 1725 W Harrison St, Chicago, IL 60612.

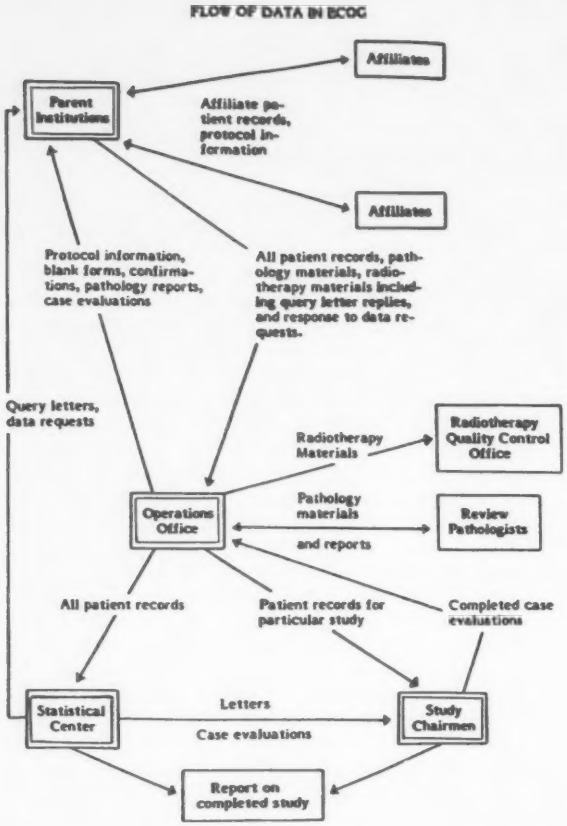


FIGURE 1.—Schematic showing how materials are handled and how institutions are involved in follow-up.

Of major importance to the quality of the data is their timeliness. The Statistical Center produces both a monthly request for outstanding data and a semiannual delinquency monitoring which looks at institutional performance in timeliness of data submission. Both of these are done by running computer programs on the Statistical Center data base. The requests detail which forms are due for a patient on a particular study. Thus, a study which is very active may be requested every 3 months, while a less active study may be requested only every 6 months. The request is run to provide reminders to the data managers at the institutions. There is no punitive action associated with this request. It also acts as a cross-check to the accuracy of the Statistical Center data base because the institutions are asked to notify the Statistical Center of any inaccuracies that they find in the request.

The delinquency monitoring which is run twice a year does lead to suspension of randomization privileges if an institution does not meet the standards set by the ECOG Executive Committee, which insists on < 20% delinquency. Table 1 reveals the impressive improvement which occurred in timeliness when these sanctions were put in place, and we continued to have < 3% delinquency on all forms, both on-study and follow-up.

The Statistical Center does not consider a case to be final and ready for major analysis until the study chairman has completed and returned each case evaluation form. Unacceptably slow turnaround times for a study chairman have been addressed and as the semiannual delinquency monitoring is carried out, a monitoring exercise is also carried out on the timely return of case evaluations by study chairmen. Problems are discussed with the Executive Officer and if necessary with the Executive Committee. The Statistical Center data manager also checks each case, and Table 2 documents the agreement

Form No. 101
Return to: ECOG Operations Office
Medical Science Center, Room 4765
420 N. Charter Street
Madison, WI 53706

Date sent 8/2/84

To: Dr. John Doe (Principal Investigator) Doe Hospital (Institution)

Patient's Name Doe, Jane Date entered on step 7 / 6 / 83

Protocol No. E9999 Phase or Step 1 Protocol Case No. 99999

Form(s) in Question Flowsheets (treatment administration)

PROBLEMS

It is unclear from the flowsheet whether or not patient received Day 8 treatment of Cycle 6. Please clarify.

Thank you,
H. Smith

REPLY

Patient requested that Day 8 Cycle 6 treatment be held as she was performing that night as "Amahl and the Night Visitors".

M. Jones

PLEASE REPLY PROMPTLY.

LEAVE BLANK

Investigator: Please keep last copy and send remaining copies to ECOG Operations Office. Revised 3/80

FIGURE 2.—Query letter.

between data manager and study chairman in 275 cases. Dr. John Bennett, the Chairman of the ECOG New Agents and Pilot Studies Committee, found similar rates of agreement in a review of pilot study data.

On-Site Audits

The second part of quality control involves on-site auditing of protocol cases. This has been practiced in the ECOG since 1978 and a total of 222 audits have been done to date, 44 audits of main institutions (some twice and a few three times) and 177 audits of satellites. This amounts to a total of approximately 2500 charts reviewed.

After several sample audits in 1978, it became apparent that guidelines for audits would have to be set and forms developed to ensure that all important items are checked and that there is uniformity of inspection from one institution to another. It was established at that time that the audit team would consist of two physicians and one data manager, and it was found that this team could examine about 20 charts in 1 day. The number of charts which could be completed was determined partly by the efficiency of the chart system in the audited institution, partly by the complexity of the protocols being looked at, and partly by the diligence with which the audited institutions had assembled materials for the procedure. It was decided that one-third of the charts would be those of an adjuvant study, one-third those of an advanced disease study in which measurements could be checked, and one-third from the complex lymphoma-leukemia group. Charts were randomly selected. When accrual to three studies was insufficient, additional studies in the same categories were added. Copies of on-study forms and flow sheets were carried from the Statistical Office to the institution by the auditors, so that it could be verified that the data in the institution agreed with what had been submitted.

Case review sheets were developed with boxes to be checked when each of the requirements had been noted. These case review sheets consist of one page on both sides, and allow for room for comment as well as check

```
SUMMARY PRODUCED: 8/ 8/84                                          PAGE 1

PROTOCOL:       PT.#:       NAME:                  DATE RAND:12/26/79
  INST :        SATELLITE :
  TREATMENT:C   STEP:INDUC
-------------------------------------------------------------------------

DATE TREATMENT STARTED:12/26/79
  DATE TREATMENT ENDED: 3/19/84
        DATE OFF-STUDY: 3/19/84


TOXICITY GRADES:                    OVERALL TUMOR RESPONSE:
----------------                    -----------------------
6=DUE TO DISEASE,9=UNK,0=NONE)

          GI:0                      OBJECTIVE RESPONSE: PR
    VOMITING:1                       DATE OF OBJECTIVE RESPONSE: 1/12/81
    DIARRHEA:0                       DATE PARTIAL RESPONSE: 1/12/81
   INFECTION:0                       STATUS OF RESPONSE:RELAPSE
    BLEEDING:0                       DATE LAST KNOWN IN
SKIN,MUC MEMB:0                        RESP/DATE RELAPSE: 3/20/84
  NEUROLOGIC:1
 RESPIRATORY:0
          GU:0                      CURRENT SURVIVAL STATUS: ALIVE
       BLOOD:0                             DATE OF STATUS: 3/20/84
       LIVER:0
       OTHER:1  >ONE TOX.


OTHER TOXICITY:HOT FLASHES  LEG EDEMA
```

FIGURE 3.—Toxicity report.

marks. The following are areas examined on all charts:

Eligibility.—In this section, age, sex, prior treatment considerations, measurability or evaluability requirements, organ function, hematologic values, performance status, and pathological confirmation of diagnosis are verified.

Response.—Auditors are required to check initial sites of involvement as claimed on the on-study and measurement forms, document the recording of the measured lesions in the chart, and make sure that response is documented in the chart on all sites being followed. Measurements appearing on ECOG measurement forms have to be supported by notes in the chart. It is also noted whether the measurements are done according to the dates that appear on the measurement sheets and the flow sheets.

Toxicity.—All toxicity claimed on ECOG summary forms is to be investigated, with flow sheet laboratory values supported in the hospital records and all toxic effects mentioned in the hospital records subsequently reported to ECOG.

Treatment.—This involves the checking of all drug doses listed on ECOG flow sheets after a recalculation of the patient's body surface area from height and weight discovered in the chart. Treatment doses and dates are checked for accuracy of recording on the flow sheets. Protocol compliance is also looked at but the Statistical Center is primarily responsible for this, as the computer program monitors compliance as flow sheets are received.

Multimodal studies.—The presence of a narrative operative report, narrative pathology report, and an indication that the patient actually went to radiation therapy are checked.

General section.—This section checks the consent form signed by both patient and physician, permits comments as to whether records were well-organized and easy to follow, and allows for recording of any other missing information (eg, estrogen and progesterone receptor values). In the past few years, we have also checked IRB approval of the protocol in the institution.

From the very first audit it became apparent that certain things were universally difficult to confirm. First,

TABLE 1.—Delinquency results

Year	No. of study forms outstanding	No. of follow-up forms outstanding
1977	310 (36%)	1198 (23%)
1980	13 (2%)	190 (5%)
1984	26 (3%)	150 (3%)

TABLE 2.—Data manager/study chairman agreement rates (275 cases)*

Category	Agreement rate (%)
Eligibility	97.5
Protocol compliance	96.7
Toxicity evaluation	98.2
Response evaluation	96.0

* Dr. Bennett found similar rates in a review of pilot study data.

informed consents were kept in a number of different places and sometimes could not be found at all. Sometimes when they were found they were not signed by the physician or witnessed. Second, expected toxic effects such as nausea from cisplatin or neuropathy from vincristine were frequently under-reported, although unusual toxic effects were well-documented. Finally, it was difficult to confirm performance status since there was rarely a description of the patient and his activities as he entered the study. The first difficulty was obviated by insisting that institutions keep copies of informed consents in a separate folder so that at the time of any audit they are instantly available. Some under-reporting of expected toxic effects continues, but for the most part, the process of the audit has improved greatly the reporting of all circumstances, both usual and unusual. It is still difficult to confirm performance status, although that, too, is improving.

Of paramount importance in an overview of the audit process is the fact that it has been apparent from the beginning that data keeping is a global process rather than something that varies from case to case. Those institutions with sloppy records, primary data slips pasted on multiple pages of the chart, scribbled notes in the margins, etc, have that problem on all charts, and those institutions whose data are timely and well-organized usually demonstrate that throughout the whole process.

Of equal importance is the fact that both auditors and auditees learn a great deal from each encounter and, whereas initially only principal investigators were invited to go on audits, the educational value was so striking that we now use all senior investigators and hope that all data managers can be included in audits in the future.

Several serious circumstances were discovered by the audits, instances in which it appeared that data had been tampered with in two satellite institutions. These two institutions were immediately eliminated from the ECOG roster and data from their charts were expunged from the computer. With those exceptions, audit problems have consisted primarily of sloppiness and the discovery that there will always be a few irretrievable data items. Those institutions which have been audited twice showed remarkable improvement between the first audit and the second, the process itself having served to tighten quality control.

Of interest (but to no clinical investigator's surprise), a certain amount of what appeared to be noncompliance turned out to be confusion over what the protocol actually required, and in this way, also, the audits have served to point up the necessity of most careful scientific and grammatical exposition of requirements and modifications.

During the last 4 years, the ECOG has been checking on drug handling, and the National Cancer Institute now requests that we check drug logs as well as inventories and the satisfactory condition of storage places. Additional forms have been developed for this and at least one of the auditors spends a part of the day carrying out this additional requirement.

With a separate sheet generated from each chart review and an overall summary prepared from those reviews, results are checked by the Group Chairman, the Audit Committee Chairman, and the Chief Data Manager. The summary is sent to the National Cancer Institute. Although there have not yet been formulated hard and fast rules as to what to do about various situations, minor aberrations are usually simply pointed out to the audited principal investigator in an "exit interview." A number of items of missing data or inaccuracies discovered lead to a request for an item-by-item response by the principal investigator, and in some instances to a repeat audit. Those institutions whose data are all there but extremely difficult to follow or "sloppy" are requested to send their data managers to a workshop or perhaps to some institution where data management is excellent for observation and additional training.

Plans for the future include continuing auditing of approximately one-third of ECOG institutions each year with a broadening base of auditors because of the educational value. Some attempt at quantification of the various problems found shall be made, although in overview deficiencies occur in $< 5\%$ of the charts. It is our conclusion that our audits have revealed data that are verifiable and investigators that are honest.

Clinical Quality Assurance in the Pharmaceutical Industry

John J. Donahue[1,*]

My objective today is to provide you with an overview of quality assurance procedures in the pharmaceutical industry. Specifically, I will concentrate on the area of clinical study quality assurance.

The data presented this morning are a combination of data derived from personal conversations with colleagues involved in quality assurance in the industry as well as a survey of 27 companies conducted by a major pharmaceutical corporation in April of this year.

Based on the data gathered in both of these studies, I would like to address the following questions and topics:

1. Why audit clinical studies?
2. How many companies are auditing, how many people are involved, and how many studies are audited annually?
3. How are studies selected for audit?
4. Types of audits, ie, study-specific versus process-oriented, what records are audited?
5. What functions are performed by the clinical quality assurance group?
6. Are certain records more important than others?
7. What percent of the data is verified?
8. What are some of the typical study problems encountered?
9. What are the advantages of the audit program to a corporation?

Why Audit Clinical Studies?

Why should clinical studies be audited at all? After all, monitors routinely check the study records. To answer this question, I would like to briefly explore the role of the study monitor in clinical trials. Study monitors are responsible for monitoring the activities of a large number of very diverse studies. While part of their charge is to assure that the administrative aspects of a study are followed by an investigator, they are under great pressure to meet corporate deadlines with respect to costs and the number of patients being completed. In an effort to meet deadlines, the administrative record-keeping aspects of a study may slide. Protocol deviations concerning patient entry may be allowed. The monitor often finds himself in the role of "goodwill" ambassador to the busy investigator who is also doing similar studies for other competing companies. Keeping the investigator happy and producing completed patient case report forms becomes the goal and providing assurances related to the record-keeping and documentation aspects of the study which must stand the test of time from a regulatory perspective may become secondary goals.

We have all heard the horror stories about fraudulent practices in clinical trials. There was the case of a Virginia clinical investigator whose office was vandalized the night before an FDA inspection; all of his records were destroyed in a whirlpool bath. The night before he was rescheduled for another audit a fire broke out in his office; and, finally, he was assaulted by unknown individuals with a paper weight the night before another rescheduled audit (1).

There was also the case of a California investigator who lost all of his data when they fell over the side of his row boat. The courts didn't believe him and he was convicted of felony fraud. It turns out that this investigator who was licensed to practice medicine in two states never attended medical school. Many quality assurance groups were started as a result of these practices; however, I would not be so naive to believe that an auditing program will catch or even totally prevent fraudulent practices in the future. At best, an audit program can test the systems and control points which a corporate sponsor utilizes for the weaknesses that might allow fraud to occur.

Auditing clinical trials has made clinical research management more aware of its regulatory and scientific responsibilities to document all aspects of clinical research. Given the time element involved in drug development, the data generated must withstand the test of time. Even after a product is marketed, there may be continuing questions or assaults on those data from regulatory agencies, patent claims, and litigants in liability cases as recent history has demonstrated.

[1]Smith Kline and French Laboratories, Philadelphia, PA.
[*]*Reprint requests to:* John J. Donahue, PhD, Smith Kline and French Laboratories, 1500 Spring Garden St, PO Box 7929, Philadelphia, PA 19101.

How Many Companies Are Auditing, How Many Personnel Are Involved, and How Many Studies Are Audited Annually?

As early as 1979, an unpublished survey of companies indicated that approximately 75% had some form of an active audit program. A survey carried out by quality assurance personnel in a major pharmaceutical corporation earlier this year revealed that roughly 90% of the companies contacted were auditing clinical studies using various systems. In addition, this study showed that the total number of clinical quality assurance personnel in a company ranged from one to ten, with a mean of three.

It was difficult to determine the average number of audits done annually due to variations in audit procedures. As we will see later, many companies do study specific audits or process audits only, while some do both. Thus, the basis for an annual audit figure could be misleading.

How Are Studies Selected for Audit?

Three major criteria for selecting a study for audit were identified by the companies surveyed. These include:

1. Whether a study is pivotal to the clinical data presented to the Food and Drug Administration (FDA) in support of the efficacy and safety of a drug product.
2. Knowledge of an impending FDA audit of a particular clinical investigator.
3. Proximity of the new drug application filing.

Other criteria listed by companies included feedback from a monitor on problems with a specific study; requests to audit by the medical monitor; studies being conducted by new medical monitors or clinical investigators without a track history.

Since some of you in the audience may be clinical investigators, I would like to briefly discuss the second item listed. FDA investigators will usually call your office prior to an inspection and tell you what study(s) they would like to inspect. They will also try to set a mutually convenient time for the inspection. I would strongly recommend that if you have not been through an FDA inspection, or even if you have, that you contact the sponsoring company and inform them of the impending inspection. You might even ask that quality assurance personnel be sent in prior to FDA's arrival to review study documents and brief your staff on what information FDA is or is not entitled to. Often, FDA investigators may ask for documents or information which they are not legally entitled to. This can lead to lengthy visits by the agency investigator and to a line of questioning that leads nowhere.

Types of Audits: Study-Specific Versus Process-Oriented, What Records Are Audited?

Roughly 70% of the companies contacted were doing study-specific audits. Usually, this type of audit involves checking all records associated with a specific study and investigator. A review of pertinent records in the sponsor archives and at the study site is done. The records reviewed may include but are not limited to the following:

1. *Protocol and Amendments*: Files should not contain outdated or draft protocols. All amendments should be present.
2. *FD 1572/1573*: Properly filled out and signed by the principal investigator.
3. *IRB Approval*: Has the study been approved by an IRB which meets all requirements of 21 CFR Part 56 (Standards for Institutional Review Boards)?
4. *Drug Accountability Records*: Does the drug disposition record agree with case report forms? For completed studies is the amount of drug shipped equal to that returned by the investigator plus that used by patients?
5. *Consent Forms*: Are all requirements of 21 CFR Part 50 met? Is there a signed form for each patient and was it signed prior to study entry?
6. *Monitoring Reports/Logs*: Were monitoring reports filled out properly and corrective actions taken?
7. *Continuing IRB Review and Approval*: Was the IRB kept up to date by the investigator and did it provide reapproval of the study at intervals required by 21 CFR Part 56?
8. *Case Report Forms/Source Records*: Do the case report forms accurately reflect the "raw data" in patient source records?

Usually, feedback from this type of audit is given either directly to the investigator or to the medical monitor, who then briefs the investigator and initiates any corrective actions deemed necessary.

Process-oriented audits are being conducted by about 30% of the companies contacted. Process-oriented audits evaluate a specific procedure such as patient consent or drug accountability compliance over a given number of studies. Usually, the studies chosen for a process audit are being done under different monitors so that a representative sampling of the monitoring staff is also selected. While process audits do not give assurances about specific studies, they do provide management with assurances that certain critical procedures needed for the success of clinical trials are being executed properly by the monitoring staff.

What Functions Are Performed by the Clinical Quality Assurance Group?

Numerous activities performed by quality assurance

groups were identified by the companies polled. The top ten activities performed by most companies include the following:

1. In-house investigator file reviews
2. Liaison for in-house FDA inspections
3. Case report form vs protocols for compliance
4. Protocol review prior to study initiation
5. Onsite investigator audits
6. Internal documentation management audits (data handling/filing)
7. Consent form review prior to study initiation
8. Clinical supply inspections
9. Training of monitoring personnel
10. Review of SOPs for medical personnel

It is interesting to note that some companies were doing onsite audits of studies done in foreign countries as well as audits of case report forms versus computer data bases and biostatistics reports.

Are Certain Records More Important Than Others? What Percent of the Data Is Verified?

While all companies surveyed agreed that all of the records cited earlier were important, they also agreed that the real bottom line question was, do the data in the case report form match those in the patient records? Most companies indicated that 20%–25% of patient case report forms might be checked against source documents during a routine audit. Usually, 100% of the data in each case report form was verified. Data include demographics, entrance criteria, concomitant/adjunctive medications, laboratory data, adverse reaction data, efficacy data, and special tests such as x-rays or ECGs.

Based on the findings from this sample, the data audit might be terminated at this point or additional case report forms could be evaluated. No company could give specific error acceptance levels for case report forms. It would appear that the auditor uses his or her discretion as to whether additional data need to be reviewed beyond the original 25%.

What Are Some of the Typical Study Problems Encountered?

Major study problems encountered by the clinical quality assurance groups interrogated include the following:

1. Patients entered into the study do not meet entrance criteria
2. Concomitant therapy not listed on case report forms
3. Drug accountability documentation incomplete; tablet counts inaccurate; outdated drug supplies being used
4. IRB documentation and correspondence between the IRB and investigator are incomplete
5. Protocol violations

What Have Been the Advantages of the Audit Program to a Corporation?

In answering this question, many respondents stated that the major advantages to the corporation were that audits had made the monitoring staff more aware of good data and documentation management practices. It also resulted in a strengthening of internal and external controls on clinical trials.

Many indicated that they had received favorable feedback from investigators who had had FDA inspections after a quality assurance group had briefed them on what to expect. These investigators felt that the inspections went well because they knew what to expect from the FDA investigator in terms of record requests and questions.

REFERENCE

1. Corporate Crime in the Pharmaceutical Industry (Braithwaite J, ed). London, Routledge and Kegan Paul, 1984.

Quality Assurance of Radiotherapy in Clinical Trials[1]

Arvin S. Glicksman,* Lawrence E. Reinstein, and Frances Laurie[2]

Multicenter cooperative group clinical trials require a uniformity of the selection, treatment, and evaluation of patients entered in all institutions. Quality assurance programs ensure that interinstitutional differences will not dilute the statistical assessment of these clinical trials. With the development of multimodality studies, the assessment of the appropriateness of treatment regimens delivered in each institution for each of the disciplines may become a much more complex problem, requiring quality assurance programs for each modality.

In the United States the Cancer and Leukemia Group B (CALGB) Radiotherapy Committee pioneered the development of quality assurance programs in radiotherapy group trials. This program has been expanded to encompass eight national and international groups in North America and Europe and is administered through the Quality Assurance Review Center (QARC) at the Rhode Island Hospital. Cumulatively, the QARC has reviewed 6200 patients on 58 protocols and is now performing approximately 2000 reviews/year on patients entered in 35 individual protocols.

We shall review how the QARC operates and show a few examples of the impact and results we have had in various groups.

Quality Assurance Process (refs 1,2)

The initial step was to ask all institutions participating in any of the groups monitored by the QARC to complete a facilities inventory to identify the available staff and equipment to perform protocols with radiotherapy components. In this way, we identified the responsible radiotherapists, medical physicists, dosimetrists, the photon teletherapy equipment used, the electron-beam capabilities, the simulation equipment available, and the computer treatment planning system at each institution. We now have in our computer file the beam characteristics of 1000 therapy apparatuses in 350 institutions. Each of these machines has been calibrated by the Radiological Physics Center (RPC) in Houston, TX, or one of the six Centers for Radiological Physics (CRP). This guarantees the uniformity of the rad throughout all the participating centers. The RPC and CRP monitor this calibration periodically by mailing thermoluscent dosimeters to each center following the initial visit. It is now possible for a physicist at the QARC to reconstruct isodose distributions and dose calculations on each protocol entry.

For a quality assurance program to work effectively, everyone involved must know what is required and what is to be done to each patient in each study. The research protocol must clearly define: (*a*) objectives of the study; (*b*) population to be studied with clear criteria of eligibility and exclusion; (*c*) clinical stage of disease allowed in the study; (*d*) patient's pathology with review by a panel of pathologists; (*e*) criteria of response; and (*f*) therapeutic regimen to be tested. The "Anatomy of a Radiotherapy Protocol (Reinstein LE, Laurie F, and Glicksman AS, manuscript submitted for publication) has been developed by the QARC staff to address the last point. The quality of radiation, type of apparatus to be used, anatomical volume to be irradiated, fraction size and dose delivered (not only to the central axis but to off-axis points in the volume which must be calculated), and exactly where the radiotherapy fits in the therapeutic regimen are all specifically spelled out in sufficient detail so that there can be no misunderstanding of the intent and the exact requirements of the treatment program. Only if all radiotherapists and physicists in each of the participating centers uniformly understand this can there be a uniformity of treatment throughout the cooperative group. We find that we must give a great deal of attention to the minute details of writing this document, but it is well worth the effort because it provides a sound basis for the ultimate analysis of the studies performed.

The review process which has been implemented in most of the groups that the QARC monitors includes both an interventional mode and a final review assessment. The interventional mode is applicable to studies where > 2 weeks of treatment are prescribed. The purpose of this form of quality assurance is that patients in-

[1]Supported by Public Health Service grant CA-29511 from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services.

[2]Quality Assurance Review Center, Departments of Radiation Medicine and Biology Research, Rhode Island Hospital, Brown University, Providence.

**Reprint requests to:* Arvin S. Glicksman, MD, Rhode Island Hospital, 593 Eddy St, Providence, RI 02902.

advertently treated in a manner inappropriate for protocol could still have the treatment modified so that the patient would remain eligible and fully evaluable for the study in which the patient has been entered.

This review process requires that within 3 days of the start of therapy, the simulator films, portal (verification films), prechemotherapy diagnostic films, photographs of the patient in the treatment position, and on-treatment initial report form (RT-1) should be submitted to the QARC. Upon receipt of this information, films are reviewed and it is determined whether they are appropriate for the protocol; the physics on-treatment form is reviewed to ensure that the planned program will meet the requirements of the study. This assessment is done on the day the information is received. Should there be a problem, a telephone call to the radiotherapist (or physicist) of record is made and the case is discussed. Generally, modifications of the treatment field and/or dosimetry are initiated so that the case will conform to protocol specifications. A letter is then sent to the radiotherapist of record and the principal investigator at that institution summarizing the substance of the conversation.

Upon completion of the radiotherapy, the summary data are submitted to the QARC. These include the radiotherapy treatment record, departmental calculations for central axis dose and off-axis points of interest to the study, and isodose plots, when applicable. All data are recorded on the RT-2 form and sent to the QARC. At approximately monthly intervals, senior radiotherapists in the group join the QARC staff for a final review of each case. On the basis of this review, an RT-3 form is generated which records the evaluation of the case for field size, shielding, treatment time, daily dose, total dose, calculates a TDF,[3] evaluates the patient position, number of fields, modality, and when applicable, off-axis point calculations. A summary letter is sent to the radiotherapist and the principal investigator. Periodically, the RT-1, RT-2, and RT-3 forms are sent to the group statistical office and are used by the protocol chairman as one component of the evaluation of the case in the protocol.

Results

Data submission.—The CALGB was the first cooperative group to initiate the radiotherapy quality assurance program. In this Group, before the quality assurance program was implemented, 75% of the radiotherapy data were not available to the study chairman for analysis. In July 1977, with approximately 50% of the data available, there was approximately a 40% deviation rate. As the quality assurance program progressed, the radiotherapists in the CALGB responded enthusiastically so that by 1979, 90% of the data were available with a < 5% deviation rate (fig 1). Similar results have now been seen in the Pediatric Oncology Group (fig 2), whose quality assurance program is 4 years old.

Physics review.—Technical deviations defined as large differences between the patient daily dose and those calculated by the QARC occur infrequently. They are usually the result of incomplete documentation of treatment planning data and poor communication within the radiotherapy department. Discrepancies ≥ 6% were found in 5.2% of the cases. Discrepancies > 10% were found in 2.4% of the calculations. The overall discrepancy between the calculated and recorded dose was 20% at the initial presentation of the radiotherapy data (fig 3). With the interventional mode of assessment, this fell to < 5% by the completion of the radiotherapy, as seen on the RT-2 form (fig 4).

In CALGB when we compared the rate of major deviations before and after the introduction of the on-treatment review process, we found a statistically significant improvement in protocol performance at the $P < .001$ level (table 1). In fact, we also examined the deviation rate at institutions before and after the QARC wrote any letters to them, whether on treatment or after treatment. Again, a highly significant improvement in the technical details of protocol participation was observed. This was clearly seen as a learning experience. We have found that with time the deviation rate for a study reviewed by the QARC will fall. This was not the case in studies not reviewed by the QARC.

Beyond the impact of the physics performed as an outgrowth of the compliance with protocol, physicists within the groups have developed a program for quality control within their own departments. Thus, the standards of daily clinical radiotherapy have been improved since the technical requirements for protocol adherence were applied to all cases treated in the participating radiotherapy departments (4).

Impact of the QARC

The experience of the QARC program in the CALGB can be used to demonstrate the impact of the quality assurance program in the assessment of two leukemia protocols: one, *7411*, was in effect prior to the development of a quality assurance program (and particularly an interventional mode), and the other, *7611*, in which the full program was implemented. Both protocols included induction chemotherapy and maintenance with intensification which included CNS prophylaxis consisting of cranial irradiation, to include the entire meningeal surfaces of the intracranial content down to the level of C2. Specifically, the posterior orbit, the base of the skull, and the entire calvarium were to be included in the radiation field. A dose of 2400 rad in 12 sessions was to be delivered using parallel opposed fields preferably with the patient in the supine position. Among the participating on-

[3]Time/dose fractionation factor—a mathematically derived equation which attempts to express biologically equivalent treatment programs (3).

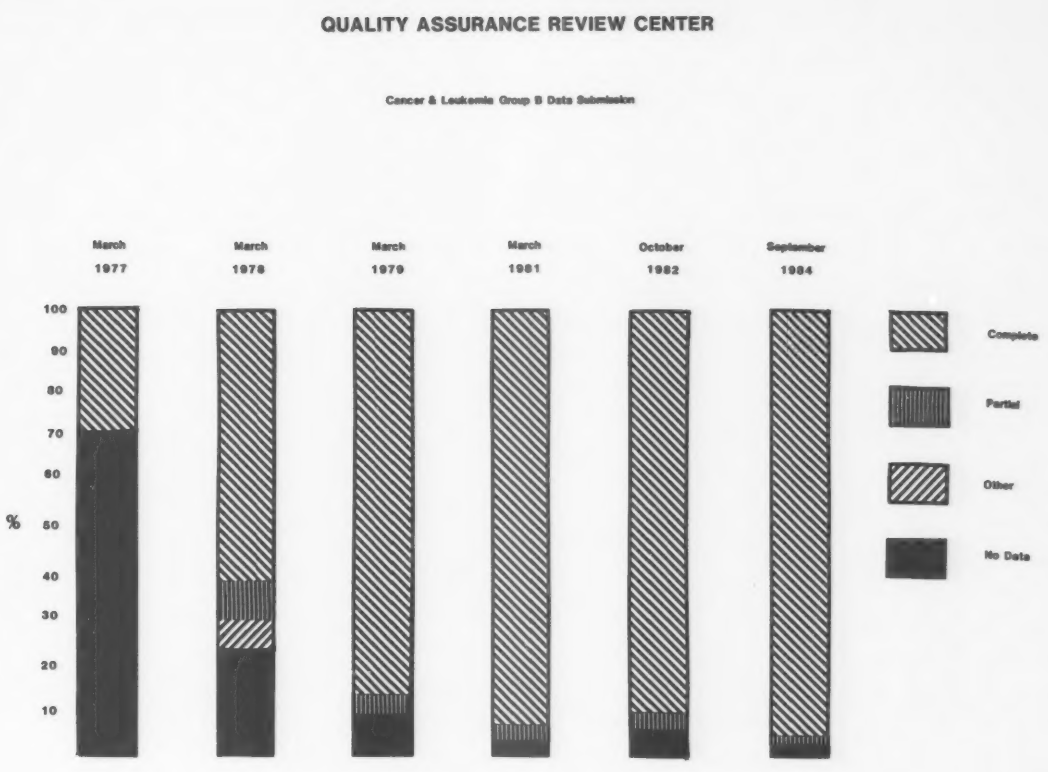


FIGURE 1.—Changes in data submission over 7-yr period of Quality Control Program for the Radiotherapy Committee of the CALGB.

cologic institutions, 46 were in the United States, four were in Switzerland, one was in Paris, and one was in Copenhagen. Twenty-seven of the 52 institutions had the status of principal centers with direct membership in the CALGB. The remaining 25 institutions had the status of satellites of the principal centers. Six hundred ten children were entered on the radiotherapy arms of these two studies.

TABLE 1.—Impact of the QARC correspondence: protocol *7611*[*]

	No. (%)—		
	Entered	Evaluable	Appropriate
Before 1st letter	171	97 (57)	37 (38)
After lst letter received	91	79 (87) $P = .001$	51 (65) $P = .001$

[*] Correspondence from the QARC in CALGB protocol *7611* affected subsequent performance in this study, both in the submission of data and the No. of cases which were appropriately treated according to the protocol. Differences before and after a letter were statistically significant at the $P = .001$ level.

An analysis of the submitted data indicated that in protocol *7411*, where no interventional quality assurance took place, only 26% of the cases entered were appropriate for the protocol in 1974, and this changed very little throughout the course of the study (fig 5). In protocol *7611*, where the full quality assurance program was operational, 37% of the cases were assessed to be appropriate for study when the protocol first opened, and the rate rose steadily to 61% by the time it was closed in 1979. The full evaluability rate in *7411* rose from 37% to 53% during the course of the study, while in *7611* it rose from 59% to 73%. This improvement was statistically significant at the $P = .001$ level and was attributed to the impact of the QARC correspondence (5,6).

Improvement in performance occurred almost entirely in the principal centers and not in the satellite institutions. This difference in performance was statistically significant at the $P = .05$ level, indicating that adherence to protocol requirements increased with increased participation in studies (table 2).

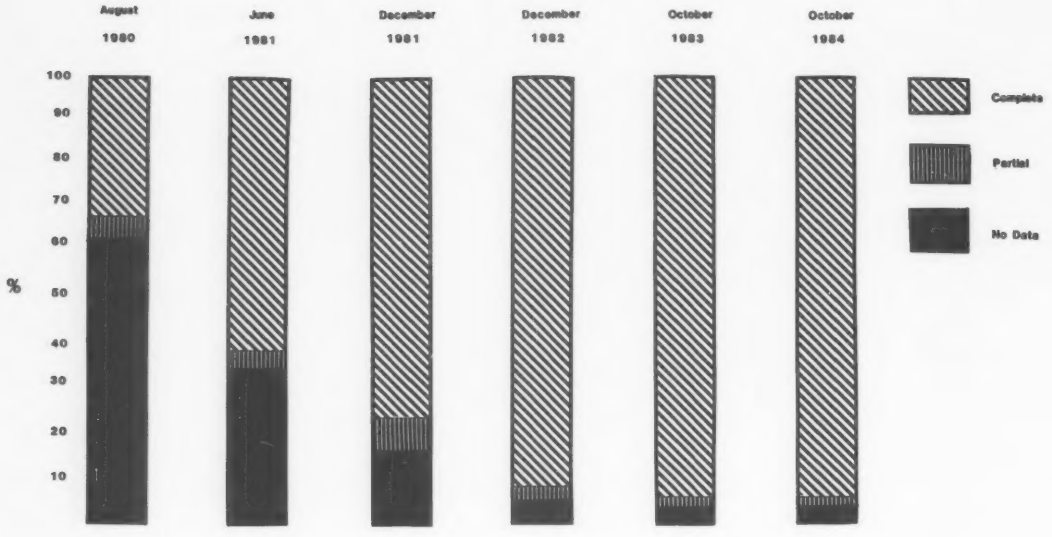


FIGURE 2.—Changes in data submission seen over 4 yrs from the newly formed Pediatric Oncology Group.

Another study in which the QARC's impact can be demonstrated is the analysis of CALGB study *7781* for limited small cell carcinoma of the lung (7). This protocol randomized patients between two regimens of chemotherapy: methotrexate, doxorubicin, lomustine, and cyclophosphamide versus cyclophosphamide, lomustine, doxorubicin, and vincristine. In both arms, radiotherapy to the primary lesion in the chest was to be delivered after 6 weeks of chemotherapy. The volume was to encompass the original, prechemotherapy tumor volume, the entire mediastinum, and both supraclavicular fossae. A tumor dose of 4500 rad was required. However, because of concern of spinal cord toxicity, particularly in the presence of chemotherapy, the spinal cord dose was reduced to 4200 rad. In their attempt to adhere to this protocol requirement, many radiotherapists inserted a posterior block at 2500–3000 rad and continued with parallel opposed fields, thus reducing the posterior mediastinal dose in most cases.

TABLE 2.—Institutional performance*

	Protocol (%)—		
	7411	*7611*	
Principal center	3/23 (13)	13/27 (48)	$P = .05$
Satellite	6/27 (22)	8/25 (32)	$P = .5$

* Institutions with data ≥ 50%. Improvement in protocol performance was predominantly at the principal centers and not at the satellite facilities. This probably related to the No. of cases entered and prior contact with the Quality Assurance Program.

A careful assessment of 105 patients, of whom 88 were evaluable, indicated that most patients did not receive full mediastinal doses although the midplane, central axis dose was appropriate in most cases. This was done by a reconstruction of the dose distribution using three-dimensional volumetric isodose techniques (8,9). There was a clear relationship between the minimum dose to the mediastinum and the percent of cases failing. When 3500–3800 rad were given, eight of ten cases had an in-field recurrence. At 3800–4000 rad, 15 of 32 cases failed

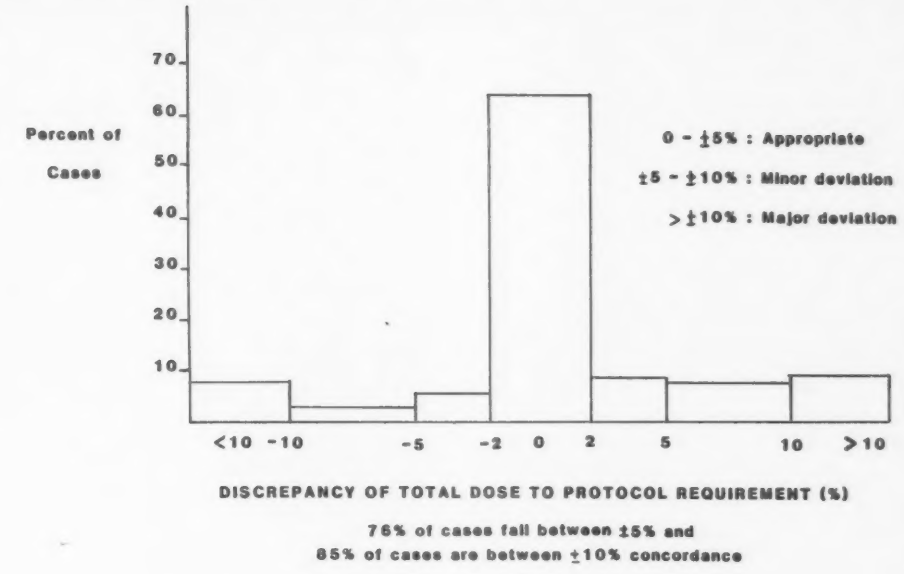


FIGURE 3.—Physics data submitted had > 15% of cases with deviations > 10% on initial review.

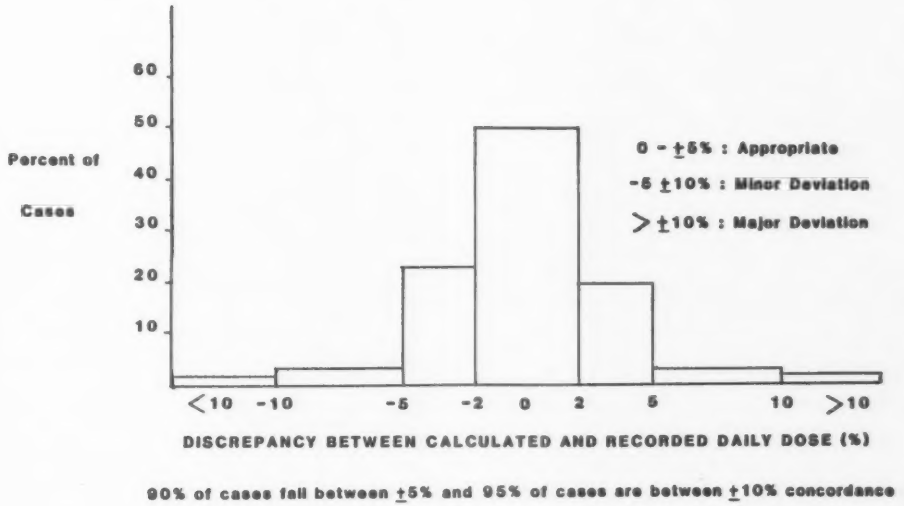


FIGURE 4.—Physics data had few technical deviations after initial review. Ninety-five percent of the cases were either appropriate or had a minor deviation.

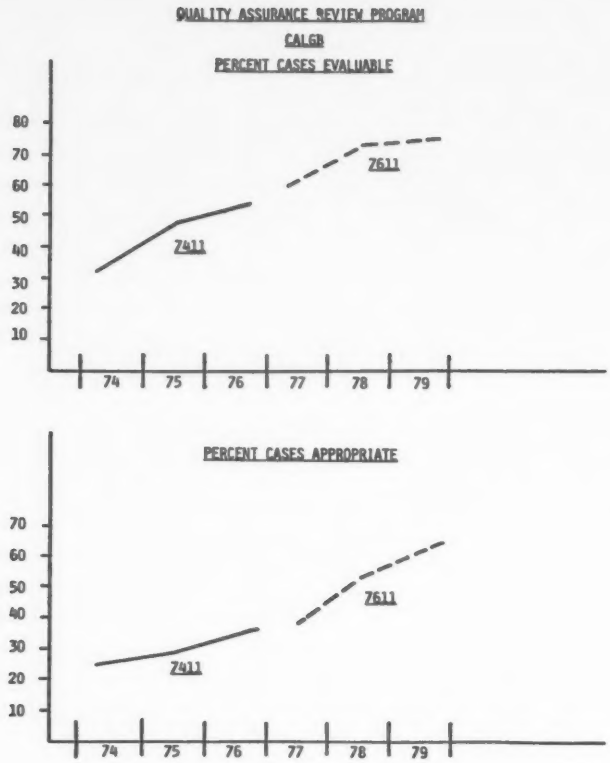


FIGURE 5.—CALGB protocol *7411* contained no quality assurance review mechanism and showed no real improvement in performance throughout the study. CALGB protocol *7611* had a quality assurance review mechanism and showed significant improvement in data submitted (evaluable cases) and the % of cases that were treated appropriately for the study.

in the chest. At 4100–4400 rad, ten of 32 patients failed. Of the 14 cases who received 4400 rad, no infield failures occurred. A linear trend test of these data had a P value of .0008 (fig 6). A dose-response curve for small cell carcinoma of the lung has been derived from these data and has been incorporated into the next study (CALGB *8083*) and subsequent small cell carcinoma of the lung protocols.

Thus, the analysis of the deviant cases had an important impact on our understanding of radiotherapy in small cell carcinoma of the lung. It had been the conventional wisdom that small cell carcinoma of the lung was a relatively radiosensitive tumor but that these cases failed because of distant disease already present at the time the patients were initially seen. Clearly, as chemotherapy has improved and patients are living longer, it is apparent that modest doses of radiation therapy are associated with local recurrences. Only as the dose is increased does local control become possible. When improved chemotherapeutic regimens become available, this information can lead to a combined approach with the prospect of cure. With intensive treatment, there may be a chance of success with this otherwise fatal illness.

CONCLUSION

The basic purpose of quality assurance programs in radiotherapy is to effectively increase the number of evaluable cases in studies so that ultimately every case entered will be completely evaluable. Radiation oncologists in multimodality study groups have recognized the importance of quality control and have responded well to the interventional review. Other disciplines in cooperative groups are in the process of developing comparable

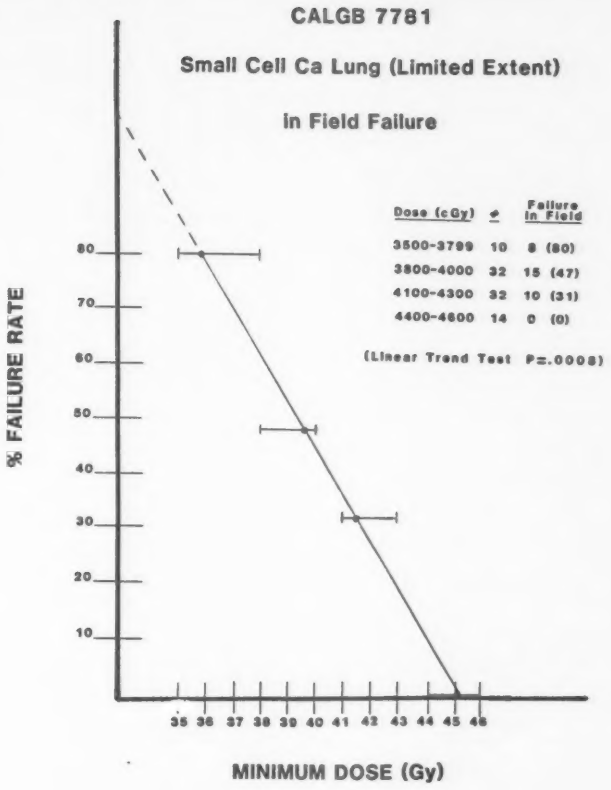


FIGURE 6.—Chest infield tumor recurrences were a function of the minimum dose to tumor and mediastinum. The central axis midplane dose was usually 4200–4500 rad. Doses were recalculated using 3-dimensional dose computational techniques.

programs which must meet the unique needs of each modality. For radiotherapy, an additional benefit has been that adherence to protocol requirements has improved the overall level of concern in the radiotherapy department with broad impact on the quality of radiotherapy performed for all patients whether in protocol studies or not.

REFERENCES

1. GLICKSMAN AS, REINSTEIN LE, BROTMAN R, ET AL. Quality assurance programs in clinical trials. Cancer Treat Rep 64:425–433, 1980.
2. GLICKSMAN AS, REINSTEIN LE, MCSHAN D, ET AL. Radiotherapy quality assurance program in a cooperative group. Int J Radiat Oncol Biol Phys 7:1561–1568, 1981.
3. ORTON CG, and ELLIS F. A simplification in the use of NSD concept in practical radiotherapy. Br J Radiol 46:529–537, 1973.
4. REINSTEIN LE. Significance and role of radiation physics in multidisciplinary tumor therapy as experienced in the CALGB. *In* Monograph in Neue Asperke Radiologische Diagnostik Und Therapie, Schweizerischen Gesellschaft fur Radiologie und Nuklear Medizin, Basel, Switzerland, 1979, pp 138–144.
5. REINSTEIN LE, LANDMANN C, ROTH J, ET AL. The effect of patient position on radiotherapy protocol deviations in the treatment of acute lymphocytic leukemia. Am J Clin Oncol 5:303–306, 1982.
6. HUENIG R, LANDMANN C, ROTH J, ET AL. Quality control of radiotherapy in acute lymphocytic leukemia protocol treatment: experience with 610 cases. Eur J Cancer Clin Oncol 19:1585–1591, 1983.
7. EATON WL, MAURER H, GLICKSMAN AS, ET AL. Relationship of infield recurrences to prescribed tumor dose in small cell carcinoma of the lung. *In* Proceedings of the 23rd Annual Meeting of the American Society for Therapeutic Radiology. Int J Radiat Oncol Biol Phys, 1981, vol 7, p 1223.
8. REINSTEIN LE, MCSHAN D, LAND RE, ET AL. Three-dimensional reconstruction of CT images for treatment planning in carcinoma of the lung. *In* Computed Tomography in Radiotherapy (Ling CC, Rogers CC, and Morton RJ, eds). New York, Raven Press, 1983, pp 155–165.
9. MCSHAN D, REINSTEIN LE, LAND RE, ET AL. Automatic contour recognition in three-dimensional treatment planning. *In* Computed Tomography in Radiotherapy (Ling CC, Rogers CC, and Morton RJ, eds). New York, Raven Press, 1983, pp 167–173.

Pathology Quality Control in the Cooperative Clinical Cancer Trial Programs

Richard L. Kempson[1,*]

Pathology quality control is an essential part of clinical cancer research. However, pathology review for every study may not be practical; therefore, review in most cooperative groups will be performed only for selected protocols. The selection process should include those neoplasms with an expected diagnostic error rate which would affect the outcome of the study. Unfortunately, complete information about the diagnostic error rate for various neoplasms is not available at the present time; however, enough information is available from the pathology review experiences of the last decade to draw some general guidelines. All lymphomas should have pathology quality control review as should rare or difficult-to-diagnose neoplasms. Early cancer is likely to cause diagnostic disagreement, so protocols designed to evaluate various treatment options for such neoplasms should include pathology review. The diagnostic error rate for far-advanced carcinoma, most adenocarcinomas of the gastrointestinal tract and squamous cell carcinomas, is so low that routine pathologic quality control review is not cost-effective, in our opinion. However, pathology review may be pathology research, and in this circumstance, pathology review will be carried out to answer research questions. Whether or not pathology quality control review is needed should be determined at the time of development of each protocol. [Cancer Treat Rep 69:1207–1210, 1985]

The purpose of pathology quality control in the Cooperative Clinical Cancer Trials Program is to ensure that all patients included in treatment result reports have had the neoplasm under investigation and that pathologic staging and sectioning requirements have been met. Pathology review is accomplished by re-examining histologic sections, cytologic smears, or other types of preparations to confirm that the morphologic and/or cytologic features in the specimen are those of the intended neoplasm. Pathologic quality control diagnoses are not necessarily used for the management of patients; rather, the review diagnosis is used to determine whether a patient is eligible for a given study. If the review interpretation differs from the initial diagnosis, the institutional pathologist must determine whether or not the review diagnosis will be the diagnosis of record used for individual patient care.

Mechanism of Pathology Quality Control Review

Pathology quality control in the cooperative groups is generally best accomplished through the mechanism of a pathology committee. This committee should have a chairman as well as rules and regulations governing review, selection of members, and other functions. Several mechanisms for slide review have evolved, depending upon the needs of the group. The most common procedure is for the slides to be sent to a central repository and then forwarded to the appropriate expert(s) for confirmation of diagnosis. The slides are then returned to the repository with a completed pathology review form. Copies of the latter are placed in the pathology committee files and in the patient's record at group headquarters. Patients whose diagnoses were not confirmed after review should *never* be included in the final analysis and evaluation of treatment results. Some cooperative groups have developed panels of experts who review slides at the time of group meetings. Others have utilized fast review by an expert or experts at a single institution. Whatever mechanism is used, it is essential that the group pathologists contact institutional pathologists to obtain their agreement to participate in the review process.

The classification of neoplasms is based on the ob-

[1]Department of Pathology, Stanford University School of Medicine, CA.

[*]*Reprint requests to:* Richard L. Kempson, MD, Laboratory of Surgical Pathology, Stanford University Medical Center, 300 Pasteur Dr, Rm P-2020, Stanford, CA 94305.

served outcome of patients whose tumors exhibit certain morphologic features. The decision as to whether a neoplasm is malignant is based primarily on nuclear changes and, to a lesser extent, on growth patterns (eg, infiltration, invasion of vessels, etc). The type of neoplasm is determined by evidence of cellular differentiation (eg, squamous cell carcinoma, rhabdomyosarcoma, etc). Expert pathologists are those who have a large experience in applying morphologic criteria for the various types of neoplasms; indeed, such experts may have developed the diagnostic criteria from their own clinical pathologic studies. Unfortunately, not all neoplasms develop morphologic features which allow easy recognition. When poorly differentiated neoplasms are encountered, diagnostic decisions are often based partially on the judgment of an expert (and hopefully agreed to by more than one expert). The pathology experts in each cooperative group should take the responsibility to inform institutional pathologists about the criteria selected for the neoplasms under study and how the various morphologic features are weighted when judgment diagnoses are made.

Sometimes, the direction of differentiation of a neoplasm will not be certain to any observer after light microscopic examination of the usual sections. For some neoplasms, ultrastructural examination, histochemical, and/or tissue immunologic stains may provide evidence of differentiation which is not discernible in ordinary sections. It is important to decide what ultrastructural and/or histochemical and immunologic features will count as evidence of differentiation prior to beginning a study. These special procedures are expensive, time-consuming, and require special handling of tissues; therefore, they should not be used for quality control unless necessary.

What Neoplasms Should Be Reviewed?

Prior to pathologists' participation in cooperative clinical trials, little was known about the error rate of pathologic diagnoses. However, even at that time it was realized that the diagnosis and classification of malignant lymphomas are often difficult. Consequently, in the 1960s when larger numbers of patients with lymphoma were entered in randomized trials, an intergroup panel of pathology experts was assembled to review lymphoma diagnoses. This panel found a significant disagreement rate between the diagnosis of the contributing pathologist and the review panel. For example, the lymphoma panel report published in 1982 noted major disagreements between the contributors and the panel pathologists concerning the diagnosis and subclassification of Hodgkin's disease in > 16% of the submitted cases (494 of 2953) (1). For non-Hodgkin's lymphoma, major disagreements were almost 28%. These results support the need for pathology quality control for lymphoma trials.

For nonlymphomatous solid neoplasms, little has been published about the accuracy of pathologic diagnoses in cooperative clinical trials, with the notable exceptions of breast carcinoma, Wilms' tumor, and childhood rhabdomyosarcoma (2–5). In our cooperative group, we found a 3%, 4%, and 8% major disagreement rate, respectively, in three studies of patients with oat cell carcinoma, ie, the diagnosis of oat cell carcinoma could not be unequivocally confirmed.[2] There was a 15% disagreement rate concerning the subclassification of oat cell carcinoma; however, in our group, and in most recently published studies, the subtype of oat cell carcinoma seems to have no bearing on the outcome. Therefore, we do not think it cost-effective to review oat cell carcinoma cases for subclassification, but we do review them to confirm the diagnosis of oat cell carcinoma because of the disagreement rate noted above. On the other hand, the diagnostic disagreement rate for non-oat cell lung carcinoma is negligible in our experience. Of course, there is a higher disagreement rate (4% in our group) for subtyping non-oat cell carcinoma. We think pathology quality control review for non-oat cell lung carcinoma protocols is indicated only if the research question is a difference in response to therapy among patients with various histologic subtypes of non-oat cell lung carcinoma.

The diagnostic disagreement rate in our group for high-grade glioma is 2% (> 300 patients), it is negligible in bladder cancer, and it is only 6% for lymphoma.[2] In our group, the diagnosis of lymphoma per se has been almost always correct, but the number of patients studied is small. The diagnosis of testicular germ cell tumors is practically always correct, as is the decision about whether the tumor is pure seminoma, nonseminoma, or a mixed germ-cell tumor. We have observed a 4% disagreement rate in subclassifying the type of nonseminomatous tumors; however, this is a diagnostic exercise which has no effect on randomization or treatment in our studies. The disagreement rate for far-advanced carcinomas, most adenocarcinomas, and squamous cell carcinomas is so low that we think routine review for quality control is unnecessary.[2]

What is our answer to the question which heads this section? We think those neoplasms for which there is a significant incidence of diagnostic and classification disagreement must be reviewed. It is obvious that a major factor in the validity of any study reporting the results of treating patients with cancer is to ensure that only patients with the neoplasms under study have been included in the treatment and control groups. For which neoplasms is there significant diagnostic disagreement? Unfortunately, for some neoplasms this is uncertain at the present time, although it is known for others. Some

[2]Kempson RL. Unpublished data from the Pathology Review Committee of the Northern California Oncology Group.

of the experience in our group has been described above. The other cooperative groups also have had experience with pathology review over the last decade, and it is probable that more information is available. Perhaps the best way to obtain this information is for the National Cancer Institute to sponsor a meeting of group pathologists at which time the data obtained from the pathology review carried out by the various cooperative groups and review panels over the last decade would be presented. There also should be participation by statisticians who could provide more information about the effect of diagnostic disagreement on the outcome of a study. With such information, pathologists would then be in a better position to decide which protocols need pathology review and which do not. This is important because pathology review is costly. There must be compensation for review pathologists, for technical personnel involved in cutting sections, and for individual institutions to forward cases to the repository. Certainly, pathology review is cost-effective for studies involving neoplasms known to cause diagnostic difficulties; performing reviews only when it will have an impact on the outcome of the study will ensure that pathology quality control review remains cost-effective.

On the basis of information available, and from our own experience, we have formulated the following guidelines for pathology review: all lymphoma studies should be reviewed, as should studies of rare or difficult-to-diagnose neoplasms such as soft tissue sarcomas, bone sarcomas, endocrine tumors, salivary gland tumors, etc. There appears to be little cost-effectiveness in quality control pathology review of far-advanced carcinomas. Most adenocarcinomas of the gastrointestinal tract do not need quality control review, with the exception of the pancreas. Our policy on lung cancer has been described above, and our results suggest quality control review of high-grade gliomas may not be cost-effective. It should be remembered that these are general guidelines, and they do not apply to every protocol. There is one point, however, on which we suspect all pathologists would agree: the earlier the cancer is detected and the more complex the classification, the more likely there is to be diagnostic disagreement. It should always be determined at the time of protocol development whether subclassification disagreements will affect the validity of a study.

Quality control in pathology can have a significant educational benefit when morphologic criteria are standardized and taught to institutional pathologists. However, there is small educational benefit, and there will be an increased need for central review, if experts use vague criteria or if they do not teach their criteria to other pathologists. It is possible that the frequency of central pathology review could be decreased if local pathologists were taught the criteria for the neoplasms under study.

Overlap Between Pathology Quality Control and Pathology Research in the Cooperative Group Setting

Pathology quality control review may also be pathology research, and the two activities are often impossible to separate. Studies of the pathology committee of the National Adjuvant Breast Cancer Project are good examples (4,5). The pathologists in this group not only documented that the patients placed on the protocols had invasive breast carcinoma, but they also recorded the size of the tumor, the histologic pattern of the neoplasm, the number of lymph nodes involved, and a large number of other features. Continued analysis of these data has uncovered valuable information about the prognosis and the indications for adjuvant therapy in patients with breast carcinoma (5). Although the diagnosis of adenocarcinoma of the large bowel is rarely ever in doubt, pathology review might be carried out to determine if the number or level of involved lymph nodes or the depth of invasion of the carcinoma has a bearing on response to therapy and outcome (6). In the Wilms' tumor project, studies correlating morphologic patterns with response to therapy and quality control were essentially intertwined and indistinguishable (2).

Role of Pathology Quality Control at the Time of Protocol Development

There should be pathology quality control input during protocol development. We think the following questions should be addressed:

1. Is pathology quality control needed for this protocol, ie, what is the chance that the neoplasm to be studied will be misdiagnosed?

The answer to this should be based on the group's past experience, the experience of other groups (unfortunately the latter is often not available, or difficult to obtain), and an expert pathologist's knowledge about the level of difficulty of diagnosing and classifying the tumor to be studied. Ideally, statistical input will also be obtained with regard to an acceptable diagnostic error rate for each protocol.

2. How fast does the review need to be done?

Review prior to randomization presents a different set of problems than pathology review during or at the completion of a study.

3. Will special procedures, such as electron microscopy, immunodiagnosis, or histochemistry be required to classify the tumors under study?

If this is so, there should be further discussion about the need for such subclassification. If it is essential for purposes of the study, then additional funds for such procedures must be allocated, and laboratories to perform the

studies will have to be identified.

4. Are the eligibility requirements based on an up-to-date pathology tumor classification?
5. Are instructions, including diagnostic criteria, available to the local pathologists?

If there will be significant alterations in the usual dissection and sectioning procedures, the institutional pathologists must be consulted, and they must agree to the proposed changes.

6. Is the proposed pathology review activity adequately funded?
7. Is a pathology review form needed?

REFERENCES

1. KIM H, and Members of the Lymphoma Panel. Pathology panel for lymphoma clinical studies: a comprehensive analysis of cases accumulated since its inception. JNCI 68:43-67, 1982.
2. BECKWITH JB. Wilms' tumor and other renal tumors of childhood: a selective review from the National Wilms' Tumor Study Pathology Center. Hum Pathol 14:481-492, 1983.
3. D'ANGIO G, EVANS A, BRESLOW N, ET AL. Management of children with Wilms' tumor: defining the risk-benefit ratio. Front Radiat Ther Oncol 16:30-39, 1981.
4. FISHER ER, GREGORIO RM, and FISHER B. The pathology of invasive breast cancer: a syllabus derived from findings of the National Surgical Adjuvant Breast Project (Protocol No. 4). Cancer 36:1-85, 1975.
5. FISHER ER, FISHER B, SASS R, ET AL. Pathologic findings from the National Surgical Adjuvant Breast Project (Protocol No. 4). XI. Bilateral breast cancer. Cancer 54:3002-3011, 1984.
6. WOLMARK N, FISHER ER, WIEAND HS, ET AL. The relationship of depth of penetration and tumor size to the number of positive nodes in Dukes C colorectal cancer. Cancer 53:2707-2712, 1984.

Prognostic Implications of Blast Cell Morphology in Childhood Acute Lymphoblastic Leukemia: A Report From the Childrens Cancer Study Group[1]

Denis R. Miller,* Mark Krailo, W. Archie Bleyer, John N. Lukens, Stuart E. Siegel, Peter R. Coccia, John Weiner, and Denman Hammond[2]

The Childrens Cancer Study Group (CCSG) has evaluated French-American-British (FAB) morphology in newly diagnosed children with acute lymphoblastic leukemia (ALL) since 1975. A modification of the FAB system was used in which individual lymphoblast cells were scored and designated as either L_1 or L_2 on the basis of distinctive morphologic characteristics. L_1 ALL was defined as < 10% L_2 cells and > 90% L_1 cells; L_2 ALL was defined as ⩾ 10% L_2 cells and < 90% L_1 cells. FAB morphology was an independent predictor of overall survival ($P = 0.02$) in CCSG-141 and a highly significant predictor of successful induction of complete remission and event-free survival in the CCSG-160 series ($P = 0.00001$). These studies involved nearly 3900 patients. Two concordance studies have been performed. In the first (1981) study, overall concordance between the FAB reference laboratory and member institutions was 76% using a two-category system (L_1, non-L_1), 86% for L_1 cases, and 47% for non-L_1 cases. In the second (1984) concordance study, the use of more stringent, semiquantitative definitions of L_1 and L_2 lymphoblasts did not improve overall (75%), L_1 (89%), or non-L_1 (46%) concordance. The results of reference laboratory classification more powerfully predicted event-free survival than did member institutions ($P = 0.016$ vs $P = 0.125$). Quality control factors (slide preparation, cellularity, staining quality, and discipline of the reviewer) did not influence concordance. These results justify the continued assignment of patients to protocols of the CCSG-100 series on the basis of the modified FAB classification. The biological significance of the FAB morphologic variants remains to be determined.
[Cancer Treat Rep 69:1211–1221, 1985]

A number of systems have been applied to classify acute lymphoblastic leukemia (ALL), but few have survived the test of time, reproducibility, and applicability (1–4). The French-American-British (FAB) (5) classification system was introduced in 1976 and identified three types of lymphoblasts: L_1, L_2, and L_3. The three subtypes were distinguished on the basis of characteristics defining the number and prominence of nucleoli, amount of cytoplasm, heterogeneity of the lymphoblast population, and the presence or absence of vacuoles. The initially proposed criteria were subjective but were considered clinically useful and reproducible by the FAB cooperative group of morphologists. However, the FAB group did not conduct a study of the prognostic importance of their classification system in childhood ALL.

The Childrens Cancer Study Group (CCSG) first used the FAB classification system in study CCSG-141, a protocol designed to evaluate prospectively a number of clinical "front-end" prognostic factors in previously untreated children with ALL (6,7). Diagnostic bone marrow slides were submitted by participating investigators to a central reference laboratory and were reviewed for FAB morphology. To account for the mixed patterns of L_1 and L_2 cells noted in approximately 15% of submitted bone marrow slides, the original FAB criteria were modified to provide for heterogeneity. In this first CCSG FAB review, five categories of ALL were identified: (*a*) "pure" L_1 ALL with ⩾ 90% L_1 cells and < 10% L_2 cells; (*b*) L_1/L_2 ALL with 89%–76% L_1 cells and 11%–24% L_2 cells; (*c*) L_2/L_1 ALL with 75%–51% L_1 cells and 25%–49% L_2 cells; (*d*) "pure" L_2

[1]Supported by Public Health Service grants (CA) from the National Cancer Institute, National Institutes of Health, Department of Health and Human Services. *See* appendix.

[2]Childrens Cancer Study Group, Los Angeles, CA.

* *Reprint requests to:* Childrens Cancer Study Group, Operations Office, 1721 Griffin Ave, Los Angeles, CA 90031.

ALL with ⩽ 50% L_1 cells and ⩾ 50% L_2 cells; and (*e*) L_3 ALL with ⩾ 25% L_3 cells.

With these modifications, the proportions of FAB subgroups in 566 patients were: L_1 = 84%; L_2 (⩽ 90% L_1 cells) = 15%; and L_3 = 1.0%. The proportions of L_1 and L_2 morphologic types were equally distributed among clinical prognostic groups, defined by age and wbc count at diagnosis.

In CCSG-141, the modified FAB classification was highly significant for predicting overall survival (P = 0.02) and of borderline significance for predicting duration of hematologic remission (P = 0.09) when the data were subjected to the Cox multivariate regression analysis (6).

Whereas some published reports from the United States and Europe (7-10) established the independent prognostic significance of FAB morphology, other investigators were unable to confirm the prognostic significance of FAB morphology. These negative reports included studies from the Pediatric Oncology Group (11), St. Jude Children's Research Hospital (12), Dana-Farber Cancer Institute (13), England (14), and Hospital St. Louis in Paris (15). Subjectivity of criteria and lack of reproducibility were cited as explaining the controversy. However, most investigators agreed that FAB subtypes L_1 and L_2 were not associated with the broad immunologic subtypes non-B, non-T, or T-cell ALL, but that L_3 FAB morphology was associated with B-cell ALL (14,16).

1981 Concordance Study (ref 17)

L_2 FAB morphology was determined to be a poor prognostic feature based upon the results of CCSG-141. In study CCSG-161, children with "good" prognostic features as defined by age and initial wbc count but with ⩾ 25% L_2 cells were assigned to the "average" prognosis study, CCSG-162.

The reproducibility of the prognostic significance of the FAB classification system when applied at an institutional level was questioned. In response to this concern, the first concordance review was done in 1981. The objectives of the study were to determine: (*a*) the concordance of FAB morphology by reference laboratory and member institutions; (*b*) the prognostic significance of the modified FAB morphologic classification; and (*c*) the feasibility of assignment of patients to subsequent treatment protocols on the basis of institutional FAB review.

Diagnostic bone marrow slides from 1113 patients entered on the CCSG-160 series were reviewed at the reference laboratory by two morphologists. Two hundred lymphoblasts were scored by the modified FAB criteria published by Miller et al (6) in 1981. Submitted slides were returned to member institutions for FAB morphology. All data were submitted to the Statistical Center for analysis of concordance, event-free survival, and disease-free survival.

The proportion of patients assigned to FAB subgroups on the basis of percent L_1 or L_2 cells was similar at the reference laboratory and at member institutions (table 1).

The overall concordance was 71.8%. Concordance of results for L_1 ALL was 86%. Reference laboratory - member institution concordance for non-L_1 ALL (⩾ 10% L_2 cells) was 47%. When L_2 ALL was defined as ⩾ 25% L_2 cells, concordance increased to 85.6%. Concordance within the individual non-L_1 subgroups was 23% for L_1/L_2 ALL, 19% for L_2/L_1 ALL, and 25% for "pure" L_2 ALL (table 2).

FAB morphology was a significant predictor of successful induction of remission. The complete remission rate in patients with L_1 ALL was 95.5%. The induction failure rate was 3.5%. The complete remission rate in patients with pure L_2 ALL (⩾ 50% L_2 cells) was only 80%. The induction failure rate was 16.3%. In children with L_1/L_2 or L_2/L_1 ALL, the complete remission rate was 90% and the induction failure rate was 6.5%.

FAB classification as applied by the reference laboratory and by member institutions was a highly significant predictor of event-free survival (fig 1; P = 0.00001). When all bone marrow slides interpreted as L_1 ALL by member institutions were reviewed by the reference laboratory and discordantly interpreted as having L_1 and non-L_1 ALL, two distinct prognostic groups with favorable and unfavorable event-free survival emerged (P = 0.003; not shown). When all slides

TABLE 1.—Distribution of FAB subgroups as determined by reference laboratory and member institutions

FAB subgroup	% of total cases reviewed	
	Reference laboratory	Member institution
L_1	78.4	77.5
L_1/L_2	12.7	13.2
L_2/L_1	5.6	4.3
L_2	3.3	5.0

TABLE 2.—Comparison of 1981/1984 concordance studies

FAB category	% concordance	
	1981	1984
L_1	86	89
L_1/L_2	23	27
L_2/L_1	19	9
L_2	25	17
L_1	86	89
Non-L_1	47	46
Overall 4 category (L_1, L_1/L_2, L_2/L_1, L_2)	72	67
Overall 2 category (L_1, non-L_1)	72	75

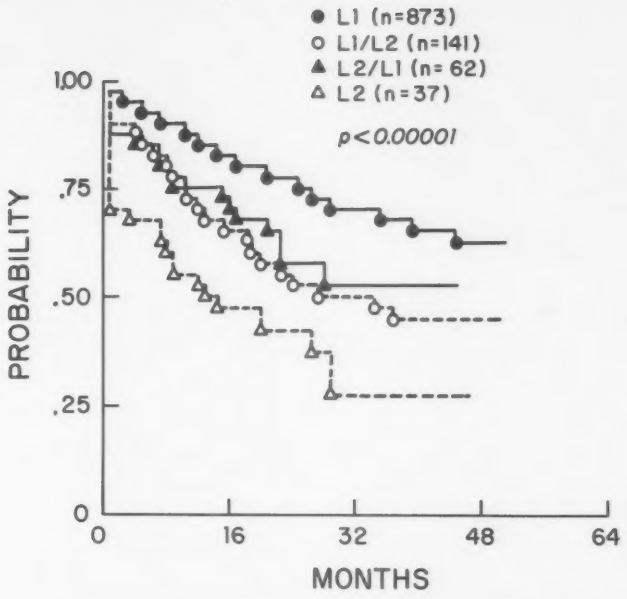


FIGURE 1.—Event-free survival of 1113 patients by FAB subgroup as coded by reference laboratory (1981 concordance study).

read by the reference laboratory as L_1 ALL were discordantly interpreted as L_1 and non-L_1 ALL by member institutions, no significant differences in the event-free survival of the two morphologic subgroups were noted ($P = 0.36$; not shown). These cross-institution/reference laboratory analyses indicated that the reference laboratory reading was more powerful in predicting event-free survival. This finding emphasizes the importance of experience and adherence to rigid criteria.

1984 Concordance Study

FAB morphologic criteria for L_1 and L_2 ALL were modified, based in part upon a more quantitative scoring system published by the FAB group in April 1981 (18). The FAB group scored the lymphoblast population on the basis of the amount of cytoplasm, prominence and number of nucleoli, regularity of the nuclear membrane, and cell size. Based upon the percentage of lymphoblasts with or without distinguishing features, the entire population was scored +2 to −4. Populations scored 0 to +2 were L_1 ALL; populations scored −1 to −4 were L_2 ALL. The reference laboratory modified this approach and scored *individual lymphoblasts* as +2 to −4 based upon the distinguishing features (table 3). A lymphoblast with a score of 0 to +2 was an L_1 lymphoblast; a score of −1 to −4 characterized a lymphoblast as L_2. This determined the percent L_1 and L_2 cells present and permitted the use of the same subgroups identified in CCSG-141 and in the 1981 concordance study. The FAB group and modified CCSG criteria are compared in table 3. These modified criteria were presented at an educational workshop on FAB morphology in June 1983, and a second concordance study using these criteria was done in 1984.

Eight slides were chosen from each institution's accrual to the CCSG-160 series. Submitted slides from the same patients but not necessarily the same slides were reviewed in the 1984 and 1981 concordance studies. Slides were selected at random, but in a manner which forced inclusion of patients' slides from all FAB subgroups (L_1, L_1/L_2, L_2/L_1, L_2) according to the original (1981) review by the reference laboratory. After review by the appropriate morphologist at the member institution, the 120 coded slides were forwarded to the reference laboratory for FAB classification. Information on variables which might affect concordance, namely, method of slide preparation, stain used, discipline of institutional reviewer, and slide quality, was obtained and analyzed. The concordance for both FAB cell percentages and FAB subtype were compared and also were related to various factors which might affect concordances.

TABLE 3.—Comparison of modified FAB group and modified CCSG criteria

Lymphoblast characteristic	Modified FAB group criteria (ref 5) (total lymphoblast population)	Modified CCSG criteria (individual lymphoblast)
Nuclear:cytoplasmic ratio	⩾ 75% high = +	High = +
	⩾ 25% low = −	Low* = −
Nucleoli	⩾ 75% 0 to 1 (small) = +	Absent or inconspicuous = +
	⩾ 25% prominent = −	Prominent ⩾ 1 = −
Nuclear membrane irregularity	⩾ 25% = −	Present = −
Cell size	⩾ 50% large = −	Large† = −
Total score	0 to 2+ = L_1 ALL	0 to 2+ = L_1 cell‡
	−1 to −4 = L_2 ALL	−1 to −4 = L_2 cell‡

*Low nuclear:cytoplasmic ratio = cytoplasm occupies ⩾ 20% of cell surface area.
†Large cell size = diameter twice that of small lymphocyte or rbc.
‡< 10% L_2 cells = L_1 ALL; ⩾ 10% L_2 cells = non-L_1 (L_2) ALL.

RESULTS

The overall concordance based upon FAB subgroups (L_1, L_1/L_2, L_2/L_1, L_2) was 67% and comparable to the 1981 concordance review of 71.8% (table 2).

In the current trial, the CCSG-100 series, patients are assigned to protocol in part on the basis of L_1 or non-L_1 FAB morphology. Overall concordance based upon the two-category system (L_1, non-L_1) was 75%. Overall concordance for L_1 cases was 89% and for non-L_1 cases was 46%, not significantly different from the results obtained in the 1981 concordance study.

The percentage of patients with non-L_1 ALL at member institutions was 5.5% greater than the non-L_1 percentage obtained in the reference laboratory. These differences reflect the further modification of the 1981 FAB criteria and the use of more stringent criteria in defining L_1 and L_2 lymphoblasts at the reference laboratory.

Quality control factors.—Seventy-seven percent (92/120) of slides were acceptable at the time of central review. Thick or dense slide preparations occurred in 10% (12/120) of submitted slides, but the presence of preparation defects in slides (thickness, poor stain, hypocellularity) or discipline of the morphologist-reviewer did not substantially affect concordance.

Prognostic Implications of 1984 Concordance Study

FAB morphology as determined by the reference laboratory remained a significant predictor ($P = 0.016$) of event-free survival despite the smaller sample size compared to the 1981 study (120 vs 1113), and use of more stringent, semiquantitative criteria to distinguish L_1 and L_2 cells (fig 2). In fact, the 1984 concordance study results were more strongly prognostic than the 1981 results. The 1984 review by member institutions did not achieve statistical significance, although the prognostic power of the 1984 institution FAB review ($P = 0.125$) was greater than the 1981 institution review ($P = 0.33$) (fig 3).

The reference laboratory classification more clearly separated L_1 and non-L_1 subgroups (fig 4). Patients rated as L_1 by the institutions were segregated into L_1 and non-L_1 subgroups by reference laboratory review. The event-free survival was worse in patients coded as non-L_1 by the reference laboratory but coded as L_1 by the member institutions ($P = 0.13$). No differences were noted in the event-free survival in children with bone marrow rated as non-L_1 by the reference laboratory but separated into L_1 and non-L_1 subgroups by the member institutions' review ($P = 0.83$) (fig 5). This indicated the greater power of the reference laboratory in classifying FAB morphology.

When the event-free survival curves of the reference laboratory review in the 1981 and 1984 studies were compared, the 1984 review more clearly separated the subgroups (fig 6). The event-free survival of patients classified as L_1 both in 1984 and in 1981 was similar to the event-free survival of patients classified as non-L_1 in 1981 but reclassified as L_1 in 1984. The event-free survival of patients classified as L_1 by the 1981 and 1984 criteria was significantly better than in patients classified as non-L_1 (L_2) in the 1981 and 1984 studies ($P = 0.03$). There were no patients classified as L_1 in the 1981 review who were reclassified as L_2 in the 1984 review. Thus, the 1984 criteria identified patients who in 1981 were classified as non-L_1 but whose event-free survival was favorable and similar to L_1 patients.

DISCUSSION

A classification system in ALL must be quantitative and reproducible, biologically significant, clinically useful, and must have both prognostic and therapeutic

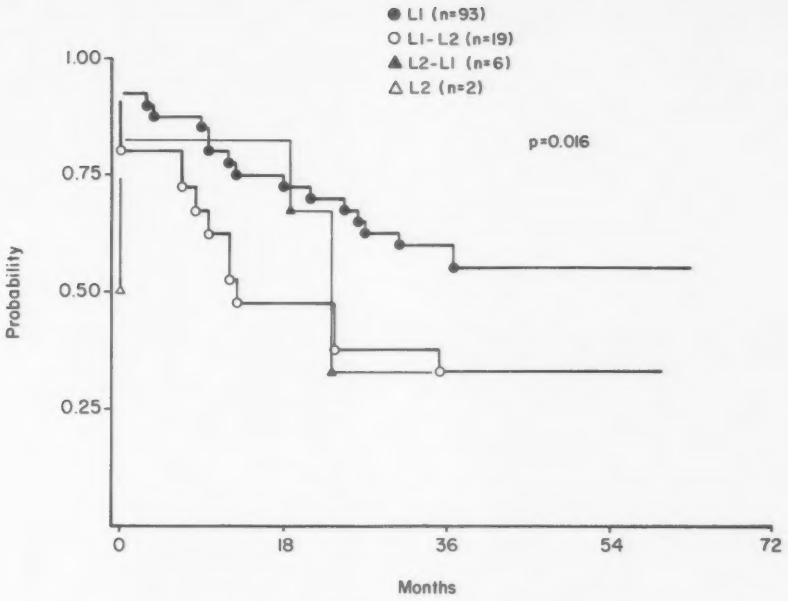


FIGURE 2.—Event-free survival of 120 patients by FAB subgroup as coded by reference laboratory (1984 concordance study).

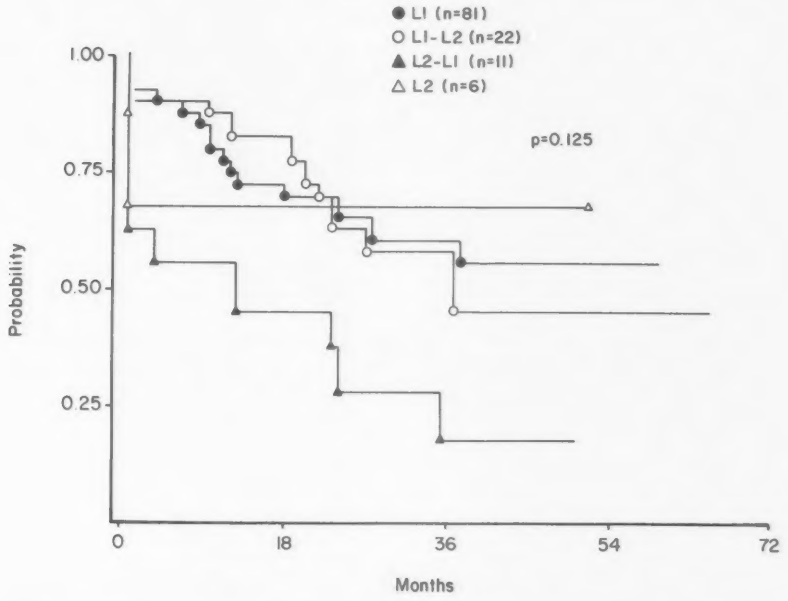


FIGURE 3.—Event-free survival by FAB subgroup as coded by member institutions in 1984 concordance study. No significant differences were noted.

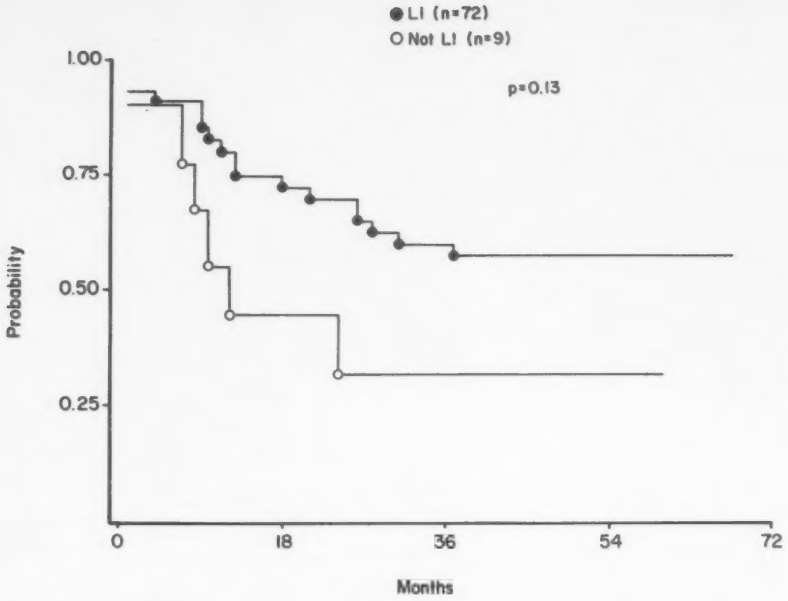


FIGURE 4.—Event-free survival of 81 patients coded as L_1 by member institutions but classified as L_1 or L_2 by reference laboratory. Patients classified as L_2 by reference laboratory but L_1 by member institutions have poorer event-free survival but difference is not statistically significant.

implications. The reproducibility of a scoring system, based upon the light microscopic evaluation of Romanowsky-stained lymphoblasts, is dependent upon several technical factors: the preparation of the bone marrow aspirate on slide or cover slip, the type and quality of stain used, and interobserver concordance. Using six different Romanowsky stains on the same sample, Tan and Lamberg (19) found a marked stain-dependent variation in morphology when lymphoblasts were evaluated for nuclear chromation and nucleoli, but found less variation in the nuclear:cytoplasmic ratio. Consistent differences in lymphoblast morphology were not observed in their admittedly small (N = 22) sample size, a finding at variance with the studies of the FAB group (5,18) and CCSG (6). As noted in our studies, only 77% of the slides were of high quality and without defects of preparation or staining (excessive thickness, hypocellularity, over- or understaining). These factors and the discipline of the morphologist did not influence the results of the concordance study.

Interobserver variability is a problem in any morphologic classification system. Most are subjective or semiquantitative at best. Mathé et al (1), using the four-category World Health Organization (WHO) classification system (microlymphoblastic, prolymphocytic, macrolymphoblastic, and prolymphoblastic), reported an overall concordance rate of 58% and an error ranging from 3% to 22% in differentiating the subgroups from one another. Other investigators, using the Mathé (WHO) system, have reported a high rate of discordance and a lack of prognostic importance (6,20–23). With training workshops and interobserver consultation, concordance improved from 63% in the original FAB study (5) to 84% in the more recent report of the modified criteria (18).

Concordance between the CCSG FAB reference laboratory and the member institutions was excellent for L_1 ALL (86% in 1981, 89% in 1984). For non-L_1 ALL, concordance was 47% in 1981 and 46% in 1984, despite educational workshops and discussion of the modified criteria. These results would imply that additional education sessions and workshops are required to improve the concordance between the reference laboratory and the member institutions. Both have improved between 1981 and 1984, but the reference laboratory has improved more. The overall two-category (L_1 vs non-L_1) concordance also changed little in the two concordance studies (72% in 1981, 75% in 1984), but the prognostic importance of the classification system was more powerful in the second study despite the smaller size (fig 6). It must be emphasized that the modified FAB criteria as published by the FAB group (18) relate

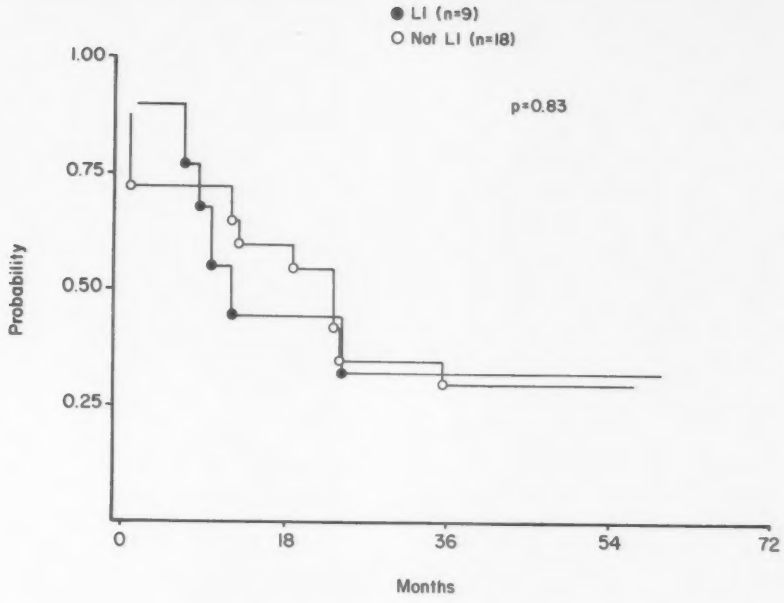


FIGURE 5.—Event-free survival of 27 patients classified as L_2 (not L_1) by reference laboratory but classified as L_1 and L_2 by member institutions. No differences were noted in the event-free survival of the FAB subgroups as determined by the member institutions.

to the entire lymphoblast population. All CCSG studies of FAB morphology are based upon the proportion of L_1 and L_2 cells. Individual cells are scored based upon the presence or absence of four distinguishing cellular characteristics: high or low nuclear:cytoplasmic ratio, absent or prominent nucleoli, absence or presence of nuclear membrane irregularity, and small or large lymphoblast size. Reported discrepancies in the prognostic significance or lack thereof of the FAB classification are most likely based upon the use of different criteria. Of importance is that CCSG alone utilizes the modified FAB criteria applied to *individual* lymphoblasts. In CCSG studies 141 and 160, FAB morphology constantly remains an important prognostic factor. In fact, in CCSG-162 ($P < 0.0001$) and CCSG-163 ($P = 0.003$), FAB morphology ranked second after the bone marrow status on Day 14 of induction as the most powerful predictor of disease-free survival. For these reasons, CCSG determined that FAB morphology would be used to assign children with ALL to current treatment protocols.

What are the implications of the concordance study vis-à-vis assignment of patients to clinical trials? Currently, patients with otherwise "good" prognostic features but with ⩾ 10% L_2 cells are entered on CCSG-105 for "average" prognosis patients, two-thirds of whom receive more intensive therapy. Children with otherwise average prognostic features but with ⩾ 10% L_2 cells are enrolled on CCSG-106, a protocol designed for "poor" prognostic patients, comparing two very aggressive regimens to a control arm of standard therapy. On the basis of the current concordance rate, what proportion of patients would be appropriately and inappropriately assigned to CCSG-100 series protocols on the basis of FAB morphology? Assuming a patient population of 1000, 85% or 850 will have L_1 morphology, and 15% or 150 will have L_2 morphology. Given an L_1 concordance rate of 89% and a non-L_1 concordance rate of 46%, 756 patients classified as L_1 will be treated as L_1 and 69 patients classified as non-L_1 will be treated as non-L_1. However, 94 L_1 patients will be classified incorrectly as non-L_1 and 81 non-L_1 patients will be classified incorrectly as L_1. Only patients with good or average prognostic features are reassigned on the basis of FAB morphology. About 15% of all patients are entered on CCSG-104 (good prognosis protocol) and about 50% are entered on CCSG-105. Patients with true L_1 morphology but misclassified as non-L_1 will be reassigned to either CCSG-105 or CCSG-106 and run the risk of overtreatment. Patients with true non-L_1 morphology but misclassified as having L_1 morphology will receive therapy on CCSG-

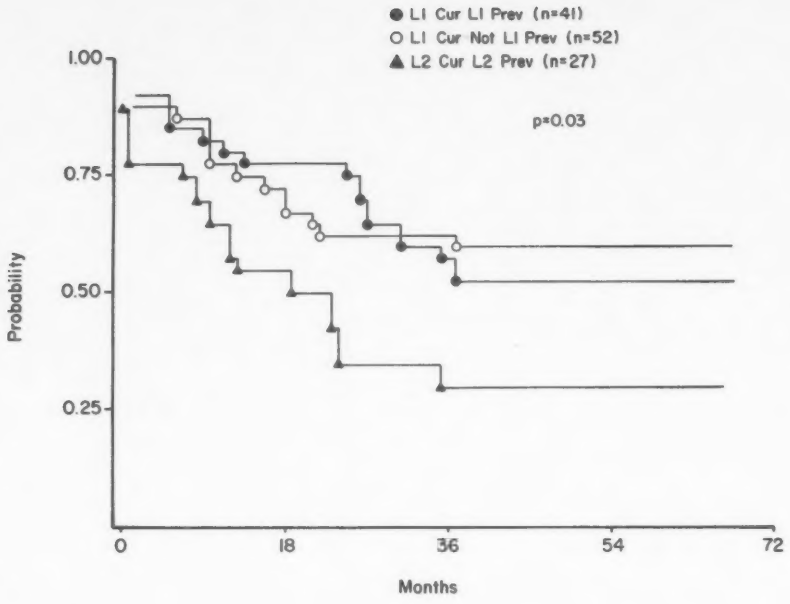


FIGURE 6.—Comparison of 1981 and 1984 concordance studies as coded by reference laboratory. Patients coded as L_1 in 1984 (Cur) and coded as either L_1 (●) or L_2 (○) in 1981 (Prev) had significantly better event-free survival than patients with L_2 morphology in both 1981 (Prev) and 1984 (Cur) (▲).

104 or CCSG-105 and run the risk of undertreatment. Tables 4 and 5 demonstrate the net effect of misclassification on assignment to more (table 4) or less (table 5) aggressive treatment protocols. As noted, 22 of 650 patients (3.4%) will be overtreated and 18/650 patients (2.8%) will be undertreated.

Most investigators now agree that L_2 lymphoblast morphology is a poor prognostic feature in ALL when evaluated in a univariate analysis. In a recent study reported from St. Jude Children's Research Hospital (24), the significance of L_2 morphology lessened when other clinical variables (initial wbc count, sex, race) were considered. When ploidy classification (pseudodiploid karyotype) was included in the multivariate model, L_2 morphology failed to add prognostic information (P = 0.0061 vs 0.0052). The biological significance of FAB morphologic variants remains to be determined. The relationship of L_1 and L_2 ALL to ploidy and cell cycle kinetics (25), cell proliferation, cytogenetics, gene rearrangements, and cell surface marker characteristics

TABLE 4.—Effect of discordance on potential overtreatment (true L_1 misclassified as non-L_1)

Appropriate protocol	CCSG-104*	CCSG-105*
Inappropriate protocol assignment	CCSG-105*	CCSG-106*
% of total entries	15	50
No. of total entries	150	500
No. (%) of L_1	128 (85)	425 (85)
No. (%) of L_1 appropriately assigned	113 (89)	378 (89)
No. (%) of L_1 misclassified as non-L_1	15 (11)	47 (11)
No. assigned to more intensive therapy	10†	12†
% assigned to more intensive therapy	3.4%	

*CCSG-104 = good prognosis protocol; CCSG-105 = intermediate prognosis protocol; and CCSG-106 = poor prognosis protocol.

†A % of patients misassigned to CCSG-105 or CCSG-106 actually receive therapy similar to that received if not reassigned because of less intensive control arms in each protocol.

TABLE 5.—Effect of discordance on potential undertreatment (true non-L_1 misclassified as L_1)

Appropriate protocol	CCSG-105*	CCSG-106*
Inappropriate protocol assignment	CCSG-104*	CCSG-105*
% of total entries	15	50
No. of total entries	150	500
No. (%) of non-L_1	22 (15)	75 (15)
No. (%) of non-L_1 appropriately assigned	10 (46)	35 (46)
No. (%) of non-L_1 misclassified as L_1	12 (54)	40 (54)
No. not assigned to more intensive therapy	8†	10†
% not assigned to more intensive therapy	2.8%	

*CCSG-104 = good prognosis protocol; CCSG-105 = intermediate prognosis protocol; and CCSG-106 = poor prognosis protocol.

†A % of patients misassigned to CCSG-104 or CCSG-105 actually receive therapy similar to that received if not reassigned because of less intensive control arms in protocols CCSG-105 and CCSG-106.

as determined by monoclonal antibodies must be studied in detail to answer this question.

In summary, two CCSG studies (141, 160 series), dating from 1975 and involving nearly 3900 patients, showed that FAB morphology as determined by the reference laboratory is a highly significant predictor of successful induction of complete remission, event-free survival (160 series), or overall survival (CCSG-141). These results justify the assignment of patients to protocols of the CCSG-100 series on the basis of FAB classification.

APPENDIX

Principal Investigators of the Childrens Cancer Study Group

Institution	*Investigator*	*Grant No.*
Group Operations Office, University of Southern California Comprehensive Cancer Center, Los Angeles	Denman Hammond John Weiner Harland Sather Richard Sposto Mark Krailo Jonathan Buckley	CA-13539
University of Michigan Medical Center, Mott Children's Hospital, Ann Arbor	Ruth Heyn	CA-02971
University of California Medical Center, San Francisco	Arthur Ablin	CA-17829
University of Wisconsin Hospital, Madison	Nasrollah Shahidi	CA-05436
Children's Orthopedic Hospital and Medical Center, University of Washington, Seattle	Ronald Chard	CA-10382
Rainbow Babies and Children's Hospital, Case Western Reserve University, Cleveland, OH	Peter Coccia	CA-20320
Children's Hospital National Medical Center, George Washington University, Washington, DC	Sanford Leikin	CA-03888
Childrens Memorial Hospital, Northwestern University, Chicago, IL	Helen Maurer	CA-07431
Children's Hospital of Los Angeles, University of Southern California	Jorge Ortega	CA-02649
Children's Hospital of Columbus, Ohio State University	Frederick Ruymann	CA-03750
Babies Hospital, Columbia University, New York, NY	Sergio Piomelli	CA-03526

Principal Investigators of the Childrens Cancer Study Group (cont'd)

Institution	*Investigator*	*Grant No.*
Children's Hospital of Pittsburgh, University of Pittsburgh, PA	Vincent Albo	CA-36015
Vanderbilt University School of Medicine, Nashville, TN	John Lukens	CA-26270
Doernbecher Memorial Hospital for Children, University of Oregon Health Sciences Center, Portland	Robert Neerhout	CA-26044
University of Minnesota Health Sciences Center, Minneapolis	Mark Nesbit	CA-07306
University of Texas Health Sciences Center, San Antonio	Paul Zeltzer	CA-36004
Children's Hospital of Philadelphia, University of Pennsylvania	Anna Meadows	CA-11796
Memorial Sloan-Kettering Cancer Center, Cornell University, New York, NY	Peter Steinberz	—
James Whitcomb Riley Hospital for Children, Indiana University, Indianapolis	Robert Baehner	CA-13809
University of Utah Medical Center, Salt Lake City	Richard O'Brien	CA-10198
Albert Einstein College of Medicine, Montefiore Medical Center, Bronx, NY	Denis R. Miller	CA-23742
Strong Memorial Hospital, University of Rochester, NY	Harvey Cohen	CA-11174
University of British Columbia, Vancouver, Canada	Mavis Teasdale	Vancouver Foundation grant CA-29013
Children's Hospital Medical Center, University of Cincinnati, OH	Beatrice Lampkin	CA-26126
Harbor General Hospital/University of California at Los Angeles and Miller Children's Medical Center, Torrance and Long Beach	Jerry Finklestein	CA-14560
University of California Medical Center, Los Angeles	Stephen Feig	CA-27678
University of Iowa Hospitals and Clinic, Iowa City	Raymond Tannous	CA-29314
Children's Hospital of Denver, CO	David Tubergen	CA-28851
Mayo Clinic, Rochester, MN	Gerald Gilchrist	CA-28882
Izaak Walton Killam Hospital for Children, Halifax, Nova Scotia, Canada	Allan Pyesmany	Nova Scotia

REFERENCES

1. MATHÉ G, BELPOMME D, DANTCHEV D, ET AL. Search for correlations between cytologic types of therapeutic sensitivity in acute leukemia. Blood Cells 1:37-52, 1975.
2. LEE SL, and GLIDEWELL O. Cytology and survival in acute lymphatic leukemia of children. Recent Results Cancer Res 43:28-34, 1973.
3. FLANDRIN G, and BERNARD J. Cytological classification of acute leukemia: a survey of 1,400 cases. Blood Cells 1:7-15, 1975.
4. PANTAZOPOULOS N, and SINKS LF. Morphological criteria for prognostication of acute lymphoblastic leukemia. Br J Haematol 27:25-29, 1974.
5. BENNETT JM, CATOVSKY D, DANIEL MT, ET AL. French-American-British (FAB) Cooperative Group proposal for the classification of acute leukemias. Br J Haematol 33:451-458, 1976.
6. MILLER DR, LEIKIN S, ALBO V, ET AL. Prognostic importance of morphology (FAB classification) in childhood acute lymphoblastic leukemia. Br J Haematol 48:199-206, 1981.
7. PALMER MK, HANN IM, JONES PM, ET AL. A score at diagnosis for predicting length of remission in childhood acute lymphoblastic leukemia. Br J Cancer 42:841-849, 1980.
8. WAGNER VM, and BAEHNER RL. Correlation of the FAB morphologic criteria and prognosis in acute lymphoblastic leukemia of

childhood. Am J Pediatr Hematol Oncol 1:103-106, 1979.
9. VIANNA MB, MAURER HS, and FERENC C. Subclassification of acute lymphoblastic leukemia in children: analysis of the reproducibility of morphologic criteria and prognostic implications. Br J Haematol 44:383-388, 1980.
10. VELETI J, REVESZ D, and SCHULER D. Morphologic diagnosis in childhood leukemia. Br J Haematol 40:501-506, 1978.
11. PULLEN DJ, CRIST WM, FALLETTA JM, ET AL. A Pediatric Oncology Group classification protocol for acute lymphocytic leukemia (ALinC 13): immunologic phenotypes and correlation with treatment results. *In* Leukemia Research: Advances in Cell Biology and Treatment (Murphy SB, and Gilbert JR, eds). New York, Elsevier Biomedical, 1983, pp 221-239.
12. BOWMAN WP, OCHS J, PUI C-H, ET AL. New directions of St. Jude total-therapy protocols: study X for standard-risk childhood acute lymphoblastic leukemia. *In* Leukemia Research: Advances in Cell Biology and Treatment (Murphy SB, and Gilbert JR, eds). New York, Elsevier Biomedical, 1983, pp 203-211.
13. SALLAN SE, WEINSTEIN HJ, and NATHAN DG. The childhood leukemias. J Pediatr 99:676-688, 1981.
14. GREAVES MF, JANOSSY G, PETO G, ET AL. Immunologically defined subclasses of acute lymphoblastic leukemia in children: their relationship to presenting features and prognosis. Br J Haematol 48:179-197, 1981.
15. JACQUILLAT C, WEIL MF, AUCLERC MF, ET AL. Application of the study of prognostic factors to the treatment of childhood (< 20 years old) acute lymphoblastic leukemia. Protocol 08LA74. Personal communication to the author, 1982.
16. COCCIA P, MILLER DR, KERSEY JH, ET AL. Relationship of blast cell surface markers and morphology (FAB classification) in childhood acute lymphocytic leukemia. Blood 54(supp 1):192a, 1979.
17. MILLER DR, COCCIA P, MCKENZIE S, ET AL. Acute lymphoblastic leukemia (ALL): concordance and prognosis. Annual Meeting of the Pediatric Society, Washington, DC 17:239A, 1982.
18. BENNETT JM, CATOVSKY D, DANIEL M-T, ET AL. The morphological classification of acute lymphoblastic leukemia: concordance among observers and clinical correlations. Br J Haematol 47:553, 1981.
19. TAN HK, and LAMBERG JD. Diagnosis of acute leukemia: variability of morphologic criteria. Am J Clin Pathol 68:440-448, 1977.
20. GIRALT M, NAVARRO DM, BUNUEL C, ET AL. Valuation and limitations of a subclassification of lymphoblastic leukemia according to morphological criteria. Sangre 20:11-18, 1975.
21. BENNETT JM, KLEMPERER MR, and SEGEL GB. Survival prediction based on morphology of lymphoblasts. Recent Results Cancer Res 43:23-27, 1973.
22. FLANDRIN G, and DANIEL MT. Practical value of cytochemical studies for the classification of acute leukemias. Recent Results Cancer Res 43:43-56, 1973.
23. POTTER VP, MILLER DR, SORELL M, ET AL. The prognostic significance of nuclear morphology in childhood acute lymphoblastic leukemia: a comparison of W.H.O. and F.A.B. classifications. A report from Childrens Cancer Study Group. Manuscript submitted for publication, Br J Haematol.
24. KALWINSKY DK, ROBERSON P, DAHL G, ET AL. Clinical relevance of lymphoblast biological features in children with acute lymphoblastic leukemia. J Clin Oncol 3:477-484, 1985.
25. SUAREZ C, ANDREEF M, MILLER DR, ET AL. DNA and RNA determination in 111 cases of childhood acute lymphoblastic leukemia (ALL) by flow cytometry correlation of FAB classification with DNA stemline and proliferation. Br J Haematol, in press.

PANEL DISCUSSION: Quality Assurance

Panelists: Drs. Loretta Itri, Daniel Kisner, Charles Moertel, and Howard Rockette

Dr. Daniel Kisner: In the Southwest Oncology Group (SWOG) Quality Assurance Program we have taken the approach that we should attempt to learn something about how we do clinical trials. After some comments from the other panelists I can give you some ideas as to what I personally think we and the National Cancer Institute (NCI) ought to be doing in the future with regard to that learning process. Dr. Hammershaimb asked me to examine several issues. One is, What should we be doing with the program, what can we learn from it, aside from fulfilling regulatory responsibilities, and in addition to following our accuracy in response assessment? I should point out that my hypothesis before doing this little experiment on response assessments was that our response discordance rate would be roughly 20%–30%. I presumed that we would be significantly overassessing response in many disease types, but our data do not support that hypothesis. Other things that need examining will hone in on how we do our clinical trials. First, we have to quantify the results of these site visits. Some groups have had very broad experience and yet the data, with regard to objectifying that experience, are extremely slim. There are too many generalizations. We need to quantify the results of these site visits across time to see whether we are measurably improving. In the SWOG we are just beginning a second round of site visits, so I cannot say whether or not there has been a learning curve. There has been a lot of discussion about improved data from the site visits, but I have not seen any data to prove it. I certainly don't have any from the SWOG. Second, it might very well be that we can learn something about what's an appropriate data management program. The NCI, once it has quantified data, can begin asking questions like: What is the appropriate data manager/patient load ratio for clinical trials? How many data managers are necessary for an institution of a certain size? What is the relationship, if any, between funding of an institution and its performance with regard to its data management? Are those institutions funded at 50% of the recommended amount doing it any differently than those that are funded at 80%? In fact, one hypothesis might be that some institutions that are now funded at 50% and 60% of the recommended amount are at bare bones with regard to data management. Their performance may not improve with time, unless time involves money. We have looked in a very anecdotal way at institutions in our group that have performed poorly with regard to our data management and these are either the unfunded institutions or the ones that are funded at 50% or 60% of the recommended amount and have far fewer data managers per patient accrued. I'm not sure that our experience is typical but it may be something we should examine across the clinical trials program and try to determine the appropriate level of data management support. The other question that might be worth asking is: Who should be doing clinical trials research? We have really almost no significant data as to whether or not the number of patients per year accrued by a given institution implies anything in regard to the institution's performance. I know that the Eastern Cooperative Oncology Group (ECOG) has published in this area; we have looked at the quality of the clinical trials results for our cancer control institutions versus our member institutions and see no real differences. But we need to look carefully at the number of patients accrued in relation to the quality of data management. We may very well be able to ask a question and appropriately answer it as to who should be doing clinical trials research, at least in collaborative studies. I think that if we fall into the passive posture of saying that the site visit program is just a regulatory requirement, we have to recognize two things. One is that the system as designed by the cooperative groups now will not de-

tect fraud. They in no way will detect fraud with random sampling of small numbers of patients. If we were to design a system to detect fraud it would be designed very differently and would be far more expensive. So in one sense all we are doing is intermittently going into institutions, with a threat that any given sloppy or fabricated chart might be identified in a site visit. We will not systematically detect fraud in this way. I don't think we should try to. Rather than adopt a regulatory posture, we should take the opportunity to identify questions we can ask about clinical trials research and what we can learn from the money that we're spending on site visits. I believe there are things we can learn, but this approach has not been uniformly embraced by all the groups.

Dr. Robert Wittes: Dan, with respect to your last comment, the cooperative groups have, in fact, detected fraud. They've detected among the only cases of fraud in clinical trials that we know, and we shall eventually be able to estimate with reasonable accuracy the extent of fraud going on in the group program. I think what we know already is that it is going to be very low, which is good. There are rotten apples in any system, but there seem to be very few doing cancer clinical trials.

Dr. Gus Turner: Dan, you know, one case out of a clinical investigator's whole submission to study may be missed if there is only one particular patient for whom data have been fabricated. But if a clinical investigator consistently fabricates data, evidence of fraud will show up sooner or later.

Dr. Luz Hammershaimb: I would like to ask Dr. Davis to comment on the data manager-to-patient ratio at the University of Wisconsin. We keep hearing about a ratio of 100 patients to one data manager. The few cancer centers that performed very well had a much lower ratio.

Dr. Tom Davis: I think it's confusing because you have studies that go for a very short period of time, so I'm sure that you can enroll a lot of patients on short studies with one data manager; on the other hand, the requirement that we all have for interim reporting of patients on adjuvant trials makes the experienced data manager's burden tougher than that of the new data manager, because (s)he has all of the old data to follow as well as the new. I guess that a person could handle 50 minimum to 100 maximum cases if there were a mix that included a lot of easy cases. This would mean that the number might be closer to 50. I know that the requirement for a CCOP (Community Clinical Oncology Program) is to put on only 50 cases. I don't think that one person can quite adequately do that in every instance.

Dr. Hammershaimb: Dr. Moertel, would you tell us about how Mayo Clinic conducts quality assurance and what features make it successful?

Dr. Charles Moertel: There are two areas that I might talk about. One is our intramural research and the other is our role as a research base for a community oncology group. With regard to both, I think quality is clearly dependent upon the commitment of the investigators and upon the commitment of leadership within either the center or the group. If you don't have those commitments you're not going to have quality. Second, I think quality depends upon highly competent people, professionals who are leading your clinical trials endeavors and who look upon quality of clinical trial conduct as the highest order of scientific investigation. In that regard I'm very well blessed, and I think that Dr. Judith O'Fallon should be up here talking much more than I because anything we might do that's right is largely because of her efforts. Within the center if I could pick out one single thing that is important to quality it is on-site quality control of data entry. That means in our clinical research areas, where patients are being entered on protocols, we have on site a study assistant who is responsible to Dr. O'Fallon. A patient may not be entered on a clinical trial until this study assistant has gone through all of the eligibility criteria and has gone through the research records to be certain that all data are there. Only after she has done all of this and edited the record on site may the patient be started on a protocol. Nobody can do anything until she initials the record. Then we also have a second line of quality control with regard to drugs, in that our nursing staff has that responsibility in conjunction with our pharmacy. When a patient shows up to get a drug, they also have a quality control responsibility which they must exercise before the drug goes into the patient. I think this type of quality control, right at the time, is terribly important and I think you

have to organize your center so that such quality control can be effectively and efficiently conducted. Anything after the fact, I'm sure, is not going to work.

With regard to our activities in the North Central Cancer Treatment Group (NCCTG), this is a group with dominant entries made by community oncologists. We at the Mayo Clinic join with them in some studies, but over 80% of our NCCTG entries are made by community centers, not by the university centers. They range between small- and medium-size towns, some towns you perhaps never heard of (St. Cloud, Minnesota, and Grand Forks, North Dakota) to some that you perhaps have heard of (Peoria, Illinois, and Des Moines, Iowa). All of these are small institutions, but they are not tiny. They are all clinics with multidisciplinary representation on the same site. They each have a substantial case accrual potential. They have to be able to submit at least 40 cases per year on protocol; otherwise they can't be members. We do not have a raft of tiny satellites. We have selected this type of membership because we can use our quality control personnel most cost effectively within a single institution. I frankly don't think a circuit rider works. Maybe somebody can defend that, but I don't think you can get on-site, timely quality control unless you have a quality control person who is right there on site at the time. For cost effectiveness you have to have at least a reasonable case entry from each participating institution to justify payment for the quality control person.

We feel the results of our quality control efforts can best be expressed in terms of patients that are totally lost to analysis. I think that this is the bottom line with regard to whether or not a study has good quality. We have discussed before that phase III studies that show a 10% magnitude of difference can be a major contribution. If a study has more than 10% of cases lost to analysis, with the accompanying potential bias, I simply don't believe the results of the study. The grossest evidence of poor study quality is the proportion of patients entered that you are not able to use eventually in your analysis. These include patients who were initially ineligible, those who are randomized but then cancel, or those who have such gross treatment violations that there is no possible way that you could use them in the analysis, for example, the patients treated with totally different drugs than are allowed on the protocol. When our group began to conduct studies in 1978 we had a loss of patients to analysis of just over 8%. This has improved, not steadily but in plateaus, to our current rate of patient loss of about 2% in the first 6 months of this year. I would suspect that is about as good as we are going to get because there is a certain background static that you can't get rid of.

What has happened during this period of time to produce this improvement? First of all, in 1978 our funding for group activities was in the neighborhood of $50 per case entry, total. Later we got a relatively small grant that raised our total funding to about $350 per case entry. Concurrently our rate of patient loss to analysis dropped to a plateau of about 6½%. During the first 6 months of 1984 our funding came up to about $700 or $800 per case entry. Either because of or in spite of this funding we saw a further decrease in the rate of patient loss. I do think that dollars have to do with this, because dollars have to do with whether we had a critical mass of quality personnel in our statistical center and whether we could hire quality control personnel to function full-time for each of the participating group members. From our perspective there has been a direct relationship between level of funding and quality of clinical cancer research.

Another vital component of quality performance in clinical trials is the pride of the investigators in quality. Is this really important to them, or does quality play second fiddle to "innovative science"? As far as I'm concerned there isn't any innovative science that goes on within a large cooperative group. This is not an appropriate tool for innovative science. Innovative science should be conducted before studies come to the level of late phase II or phase III, where the cooperative groups have their strongest role. The cooperative group should have its major commitment to the major ingredient that produces believable results in such studies and that's high-quality performance.

At all of our NCCTG meetings a large part of the program is spent reviewing each member's quality of performance. The full record for the membership is displayed, not only major violations but minor violations. We see the members who are not getting laboratory tests on time and we ask why not. We see those who are not sending their records in on

time and we ask why not. All of this dirty laundry is put on the line and each member stands up, rather like at an Alcoholics Anonymous meeting, and he discusses his deficiencies and what he is going to do to improve. I really think it has to be a religion if you're going to get good quality control. It has to be a religion that is shared by everyone who plays the game.

Another important ingredient in quality control is timeliness. The people in our group have to have their on-study forms submitted 2 weeks after the patient enters a study. If the forms are late, there is a little black mark. We give them another 2 weeks. If the forms still aren't in, there is a big black mark. Ninety percent of our records come in within this time limit. They are edited immediately (within 7 days) for quality problems by our statistical center personnel. If major problems show up they are immediately brought to my attention for appropriate action. Within 2 weeks after the forms arrive at the statistical center they are edited and then reviewed by the Mayo physician who is responsible for quality control of that particular study. Over a period of no longer than 6 weeks following patient entry we should have completed all phases of quality control review. If problems are unearthed we promptly feed back to the individual who made the mistake. This doesn't occur 6–12 months later, when the study may have been closed and the investigator had made the same mistake a dozen times. I think this degree of timeliness and this degree of discipline is the only way you're going to come up with a reasonable rate of data loss due to protocol violations.

Another important feature of quality control within our group is the people who answer the telephones when patients are entered on study. When these people are checking eligibility criteria they have to be hard-nosed. They don't ask such general questions as: Is this patient eligible? Are his blood counts alright? Are his chemistries alright? Was his surgery conducted 3 or 4 weeks ago (if that's an important element of the protocol)? They ask: What is the wbc count and on what date was that wbc count taken? What is this patient's creatinine and on what date was that serum creatinine determined? On what date was surgery performed? I think we should make our entry procedures sufficiently specific that somebody has to be a liar to enter an ineligible patient. In this regard, the group leadership must stand behind the study assistants who are answering the telephone. We have to support them when some physician breathes heavily about getting a patient into a study who happens to be a little bit deviant from the entry criteria. These people have instructions that they give those calls to me. They say, "Sorry I can't discuss it, but we'll get Dr. Moertel on the phone and he'll talk to you about it," and I do. They know that they've got that type of support, and also I hope that they know that we are proud of the job that they are doing. I have to keep reminding myself, and Dr. O'Fallon keeps reminding me, that we have to let our quality control people know that they are important to us. We also encourage the investigators out in the field to let their own quality control people know that they are important and to pat them on the back regularly. A lot of philosophy, but I think it's this type of spirit that has to pervade if we are going to get quality control and avoid many of the very serious problems we've seen in years past.

Dr. Judith O'Fallon: We have, at all times, about 200 protocols going on that are active and over 100 protocols that are eligible for patient entry. With that big quantum I would like to just amplify what Dr. Moertel said about the importance of the study assistants, who are our key quality control personnel. If a physician has just evaluated a patient, the chart is immediately passed on to the study assistant; while the patient is being treated, this chart is edited right there, so if there is anything that is missing from the chart, such as toxicity, the study assistant will pick that up right away. Thus, there are very few missing data. In order to be able to do that we have some quality control materials. One is a highly structured oncology record. Our chart is highly structured and we make sure that everything that a protocol requires has a place and room to document it. Therefore, the study assistant is able to pick what is missing or what seems to be illogical. Another thing that we have is what we call "protocol abstracts." That is another important quality control material. If a protocol is a 20-page document, that protocol is abstracted to include all the important features about the protocol—eligibility criteria, stratification factors, randomization schema, treatment schema, treatment schedule and what is to be treated, cri-

teria for modifying the treatment, toxicity, etc. These protocol abstracts are in every patient's chart, so with over 200 protocols going on and the resulting treatment, the study assistant and the physician have some documents right there to be able to follow the protocol.

Dr. Loretta Itri: One of the questions I have been asked to address relates to the necessity for the depth and intensity of monitoring done by the pharmaceutical industry. I am pleased to be the speaker following Dr. Moertel, since it is my opinion that if all clinical trials were conducted as carefully as at the Mayo Clinic, it would obviate much of the need for the in-depth monitoring practiced by industry. Currently, we routinely perform on-site examination of records and drug supply every 4–6 weeks. At some institutions, we have provided the only quality control for clinical data. The two primary problems are verification of response data and drug accountability. Other areas of data collection which have presented significant problems include quantitation of performance status and underreporting of expected toxic reactions, especially when investigators become more accustomed to dealing with an experimental drug. Some of these problems have been quite serious, especially the issues of drug accountability and safety, and these have largely accounted for current monitoring practices of industry. Basically, the need for accurate recordkeeping and timely reporting of adverse effects is no different for industry than for academia. Therefore, our problems are also similar. We have the same difficulty verifying data which are recorded on paper and relating these data to the actual clinical event. We can, and frequently do, examine source material, such as chest x-rays, to verify a clinical response. However, we cannot easily quantify the measurements of a skin nodule, for instance, unless we are actually examining the patient. In order to become comfortable with the situation, our monitors try to work very closely with the investigators to develop a sense of mutual trust and understanding.

Another area of extreme importance to industry is the collection of survival data. This information is now generally required by the Food and Drug Administration (FDA) for the approval of new cancer drugs and is also of great importance to anyone conducting clinical trials of cancer patients. Acquisition of this sort of data is frequently difficult. Since patients are commonly referred from great distances to centers where protocols are being conducted, they are commonly lost to follow-up once they have completed treatment. Another complicating factor is the practice of entering patients onto other experimental drug trials after they have completed treatment. The effect of this practice on survival is currently unknown, and I am unsure that a realistic method for the collection of survival data suitable for our needs is available. It is pleasing to hear the need for quality control being expressed by the centers, cooperative groups, and investigators. As Dr. Moertel has stated, we must all take more pride in the correct conduct of clinical trials. This is indeed a first-rate science. We have perhaps labored too long under the perception that clinical research is less important. As the clinical community asserts itself in this regard, improved data collection will provide benefits for all sectors. If industry were convinced that data were routinely being collected coherently and cohesively, monitoring practices could be significantly altered.

Another issue which deserves some discussion is computerized, remote data entry, which may radically alter industry standards for monitoring of clinical trials. There are problems with this method, but as the technology improves, many of these will be obviated. It is clear that the method is expensive and will require intensive training of personnel for proper utilization. It will also take time for industry to become comfortable with less personal contact. Nevertheless, the increasing quantity and complexity of data suggest that this is a very reasonable alternative. In general, I believe that there has been significant progress towards the amelioration of many of the problems discussed. As Dr. Wittes said earlier, there is fertile ground for increased collaboration between industry and the cooperative groups and centers. Our goals are not always the same, but the quality of our clinical research should be, and we should settle for nothing but the best. The collection of complete and accurate data and the realization that this is a worthwhile effort by all involved parties will provide the basis for a very positive future.

Dr. Moertel: Could I ask a question with regard to what I think is one of the most diffi-

cult segments of working with pharmaceutical companies, ie, the volume of paper? I know the one study not involving your company that we've recently become involved in where the amount of information to be collected on a given patient, one patient, one entry, is presented in a book that is larger than our Rochester telephone directory. You must seriously wonder if this type of thing is not, in fact, counterproductive to high quality. Because certainly if you are going to be recording reams and reams of stuff either you are not going to be using your best quality personnel (you would use people in training because you want experienced people out doing something more useful) or if you are using your highest quality personnel to record those data, they are not perhaps where they should be, which is on the research rather than filling out the reams of paper. I know that there is not a good answer and perhaps you have addressed one of the potential good answers, which is in computerization of these data. But since that is a few years away, do you have suggestions with regard to the paperwork problem?

Dr. Itri: We try to be reasonable in the design of our case report forms. Unfortunately, it is almost impossible at the start of a clinical trial to be certain of all the information, particularly toxicity data, which should be collected. There are, of course, routine laboratory test results that are always requested, but unexpected toxic effects frequently develop, most often in late-phase, large-scale clinical testing. Industry may then face an FDA inquiry with an inadequate database. In order to avoid this, we have, in some instances, developed a habit of collecting data that ordinarily might not be required or desirable. The amelioration of this problem probably will depend upon better communication among the investigator, industry, and the FDA. Together, intelligent and pragmatic decisions must be made regarding the type of data to be collected. Unfortunately, there have been frequent communication difficulties.

Dr. Turner: The FDA does not always see the protocol before it goes out to the clinical investigators. We may see that protocol only at the end of a year's reporting time. The FDA sees only the first protocol in an IND. After that it is up to the firm to decide if it wishes to communicate with the agency to see whether the protocol is appropriate. Now it's true that you may not want to collect all the data today, and 6 months or a year down the line something shows up in the trial and then you may wish you had all those data from the year before. The firms are starting to cover themselves by getting voluminous case report forms. I have to review those case report forms just as you have to fill them out. Sometimes when I send to a firm for a copy of a study, I get three boxes of case report forms on 40 patients. If the clinical investigator thinks that something is not right with that protocol he's got the right not to do the study. Instead of trying to do a haphazard study or fill forms out inappropriately and then say, "Well I didn't really intend to do that" later on, I think it's up to the clinical investigator to take the position that that is not the study he wants to do and then walk away from it.

Dr. Hammershaimb: In relation to the question about the difficulty in following up survival, I have asked Dr. Rockette to tell us a little about how the National Surgical Adjuvant Project for Breast and Bowel Cancers (NSABP) approaches these problems, which are particularly important in adjuvant studies.

Dr. Howard Rockette: The ultimate success of any clinical trial depends upon the timely submission of complete and accurate data on follow-up. Maintaining adequate long-term follow-up is always difficult and time-consuming. Some NSABP protocols now have an average time on study exceeding 10 years and, as the interval increases, so does the probability that the patient is no longer being treated by the same physician, or that the physician or even the institution is no longer participating in our trials. There are three aspects that are part of the NSABP long-term follow-up procedures that help to maintain a reasonable follow-up.

First, we educate institutional data managers immediately in regard to both the difficulty and the importance of good follow-up. Through our national, semiannual workshops and in procedure handbooks, the NSABP presents follow-up guidelines that reflect experience gained from the surveillance of 15,000 patients. These guidelines include steps on how to avoid potential follow-up problems with health care facilities, physicians, and patients. Also provided is information on where to find patient information within their own insti-

tutions, in other institutions, and in their community.

Second, we closely monitor delinquency by generating delinquency summaries for each participating institution. Monthly listings of patients with inadequate follow-up are sent to the appropriate institution. An institution's record on follow-up is one of the parameters that is used in the determination of institutional performance. This institutional performance is assessed semiannually and is a major part of the evaluation of an institution's continued participation in NSABP trials.

Last, the NSABP recognizes a need to go to the institution and obtain information if all other measures fail. The NSABP employs a circuit-riding data coordinator who specializes in follow-up and who spends approximately 4 months of the year visiting institutions for the collection of follow-up data. This type of effort is particularly important for institutions that are no longer participating in our trials.

In discussing follow-up one must distinguish between efforts to determine recurrence of disease and efforts to measure survival. The previously discussed procedures relate to both. However, to ascertain survival there are many additional methods that can be used that have been employed in many long-term epidemiological studies. These sources include the Social Security Administration, Motor Vehicles Bureau, and National Death Registry. Using these additional sources, a greater than 95% ascertainment of survival at 10 years is not unrealistic.

Fellowship Programs From the International Union Against Cancer (UICC)

The American Cancer Society—Eleanor Roosevelt International Cancer Fellowships are granted normally for 1 year. The closing date for receiving applications is October 1.

The Yamagiwa-Yoshida Memorial International Cancer Study Grants are awarded for periods ranging from 28 to 90 days. Closing dates for receiving applications are June 30 and December 31.

The International Cancer Research Technology Transfer awards are for periods up to 28 days. Applications are reviewed on a continuous basis.

For further information, contact International Union Against Cancer, Rue du Conseil-Général 3, 1205 Geneva, Switzerland (telephone: [41-22] 20 18 11; telex: 429 724).

International Workshop on Quality of Life Assessment and Cancer Treatment

Milan, Italy, December 11–13, 1985

This workshop will be held at the Istituto Nazionale per lo Studio e la Cura dei Tumori, to acquaint oncologists interested in psychosocial oncology with the possibilities of implementing psychosocial and behavioral outcomes into cancer clinical trials. The program will be developed along two lines: (*a*) short didactic introduction into methodology and measurement in cancer clinical trials; and (*b*) application in trials.

For further information, contact: Scientific and Organizing Secretariat, Fondazione Floriani, Vicolo Fiori 2, 20121 Milan, Italy (telephone: 02-877.240/805.9070).

First International Symposium: Multimodality Treatment of Head and Neck Cancer

Genoa, Italy, December 17–18, 1985

The Istituto Nazionale per la Ricerca sul Cancro is organizing this conference as a forum for the most active and prominent investigators in cell kinetics, prognostic factors, diagnosis, staging, evaluation of response, and single-modality treatment of advanced stages of head and neck cancer. Particular emphasis will be given to combined modalities and second-line treatments.

For further information, contact: M. Merlano, MD, Istituto Nazionale per la Ricerca sul Cancro, Viale Benedetto XV, 16132 Genoa, Italy (telephone: 010-300770).

First Charles Heidelberger Pharmacology Seminar: Interface Between Cancer Pharmacology and Cell Biology

Los Angeles, California, January 13, 1986

The Comprehensive Cancer Center of the University of Southern California School of Medicine will sponsor this seminar that will cover areas of research that promise exciting new developments in cancer chemotherapy within the next decade. Topics include: biochemical modifiers; enzymes as targets for chemotherapy; methylphosphonate oligonucleotides; DNA-directed drugs; topoisomerase; gene amplification; and drug resistance.

There is open admission with no charge for attendance. Early registration is advised due to limited space.

For further information, contact: John Holcenberg, MD, c/o USC Comprehensive Cancer Center, 2025 Zonal Ave, Rm 508, Los Angeles, CA 90033.

Ninth Annual Course on Infectious Disease in Clinical Practice

Aspen, Colorado, January 25–February 1, 1986

The Extended Programs in Medical Education, University of California, San Francisco, is sponsoring this course on infectious disease in clinical practice which is scheduled to be held at the Conference Center, Snowmass at Aspen, Colorado. This course meets the criteria for 24 hours of AMA/CMA category 1 hours. Fees are: $375 for physicians, $325 for allied health, and $275 for interns and residents.

For further information, contact: Terrie Van Alen, Infectious Diseases Division, San Francisco General Hospital, 1001 Potrero Ave, 5H22, San Francisco, CA 94110 (telephone: 415-821-8200, office hours T, W, Th 8–11 AM).

The University of California, Irvine, First International Cancer Conference

Newport Beach, California, February 13–15, 1986

This 3-day conference, sponsored by the University of California, Irvine Cancer Center, the Orange County Unit of the American Cancer Society, and the Memorial Medical Center of Long Beach-University of California, Irvine, Center for Health Education, offers 22.5 hours of Category I credit, and will be held at the Newport Beach Marriott Hotel and Tennis Club.

For further information, contact: Assistant Director, Center for Health Education, 2801 Atlantic Ave, Long Beach, CA 90801 (telephone: 213-595-3823).

Counseling Strategies for Physicians

Oahu, Hawaii, March 24–29, 1986

The University of Texas at Tylor, in cooperation with The University of Texas Health Science Center at Houston, is sponsoring this meeting to introduce the latest techniques and research findings on matters that influence doctor-patient relationships. Faculty will discuss social change, family violence, addiction, and human sexuality as considerations in dealing with patients, and they will present problems likely to be encountered with specific groups of clients such as teenagers, minority groups, aging clients, and pregnant women.

The registration fee is $325; the program is approved for 17 hours in AMA/PRA Category 1 and is AAFP prescribed.

For further information, contact: Division of Continuing Education, The University of Texas Health Science Center at Houston, 1100 Holcombe, Rm 1501, PO Box 20367, Houston, TX 77225. Telephone—within Texas: 1-800-392-4900-437, Program Code 975 DLB; outside Texas: 1-800-332-8747.

1986 Annual Meeting of the American Radium Society

San Francisco, California, April 6–10, 1986

For further information, contact: Ms. Suzanne Bohn, Executive Secretary, American Radium Society, 925 Chestnut St, Philadelphia, PA 19107 (telephone: 215-574-3179).

First International Congress on Cancer and Hormones

Rome, Italy, April 23–27, 1986

This conference will include such topics as newer diagnostic and therapeutic approaches; epidemiology of endocrine-related cancer; hormone receptors and tumors; oncogenes, growth factors, and small effectors; hormones and cell proliferation; and cancers arising in endocrine glands.

For further information, contact: Stefano Iacobelli, MD, Laboratorio di Endocrinologia Molecolare, Istituto di Clinica Ostetrica e Ginecologica, Università Cattolica del S. Cuore, L.go Agostino Gemelli, 8, 00168 Rome, Italy (telephone: (06) 33.054.496; telex: 611330 UNCATRO).

Fundamental Problems in Breast Cancer International Symposium

Banff, Alberta, Canada, April 27–29, 1986

The Breast Unit of the Cross Cancer Institute and the Division of Oncology, University of Alberta, are organizing this international symposium. Topics for discussion include: natural history of clinical significance of breast cancers detected by imaging techniques; endocrinology in growth and therapy of breast cancer; overview of important developments in breast cancer research; local and systemic therapy; radiological imaging techniques; future directions in therapy of breast cancer; and problems and priorities of cancer care and research in other countries.

This 3-day symposium will be held at the Banff Springs Hotel, Banff, Canada.

For further information, contact: Secretary, Breast Unit, Cross Cancer Institute, 11560 University Ave, Edmonton, Alberta, Canada T6G 1Z2.

International Symposium on Hormonal Manipulation of Cancer: Peptides, Growth Factors, and New (Anti) Steroidal Agents

Rotterdam, The Netherlands, June 4–6, 1986

This symposium is focused on new therapeutic approaches in the area of potential hormone-dependent tumors with respect to both peptide hormones and (anti) steroidal agents. Topics to be discussed are: LHRH-analogs; somatostatin-analogs; growth factors, peptide receptors, hyperprolactinemia; rescue or recruitment studies; new antisteroidal agents; new aromatase inhibitors; new applications for hormonal treatment; and alternating hormonal treatment. A free poster session will be included.

The symposium will be held in the Congress Center "De Doelen" in Rotterdam, The Netherlands. Abstract forms must be received before March 1, 1986.

For further information, contact: Congress Secretariat/Trial- and Datadepartment, International Symposium "Hormonal Manipulation of Cancer," The Dr. Daniel den Hoed Cancer Center, PO Box 5201, 3008 AE Rotterdam, The Netherlands.

Fourth Annual European Organization for Cooperation in Cancer Prevention Studies (ECP) Symposium: Concepts and Theories in Carcinogenesis

Brugge, Belgium, June 12–13, 1986

This 3-day conference will be a combination of review lectures, discussion panels, and poster sessions. The following topics will be discussed: biochemistry of cell proliferation and differentiation; normal versus neoplastic tissue behavior, the multistep process from the normal cell to cancer; tumor/host adaptation-immune mechanisms; mechanisms in radiation carcinogenesis and chemical carcinogenesis, role of free radicals; oncogenes and RNA viruses; DNA viruses and carcinogenesis; genetics and cancer; hormones and carcinogenesis; stress, aging, and cancer risks; multifactorial aspects in carcinogenesis; and the essential versus occasional aspects of carcinogenesis.

For further information, contact: ECP Secretariat, Avenue Lambeau 62, B-1200, Brussels, Belgium (telex: 62173 PRPB).

35th National Meeting of the Italian Neurosurgical Society

Pavia, Italy, September 18–20, 1986

Topics of discussion at this meeting will include cerebral aneurisms, pathology of the cervical column, brain tumors in infancy, spinal tumors, cerebral ischemia, and pain therapy.

For further information, contact: Prof. Pietro Paoletti, Department of Surgery, Neurosurgical Clinic, University of Pavia, San Matteo Hospital, 27100 Pavia, Italy (telephone: 0382/422231).

☆U.S. GOVERNMENT PRINTING OFFICE: 1985-461-311:10013

To: CTR–CAN
SUPERINTENDENT OF DOCUMENTS
U.S. Government Printing Office
Washington, D.C. 20402

Subscription prices:

$29 per year domestic
$36.25 per year foreign

Make checks payable to the Superintendent of Documents

☐ I want to receive my own copy of the *Cancer Treatment Reports* for the next year.

☐ Here's my check for $_____ made out to the Superintendent of Documents.

Name _____

Address _____

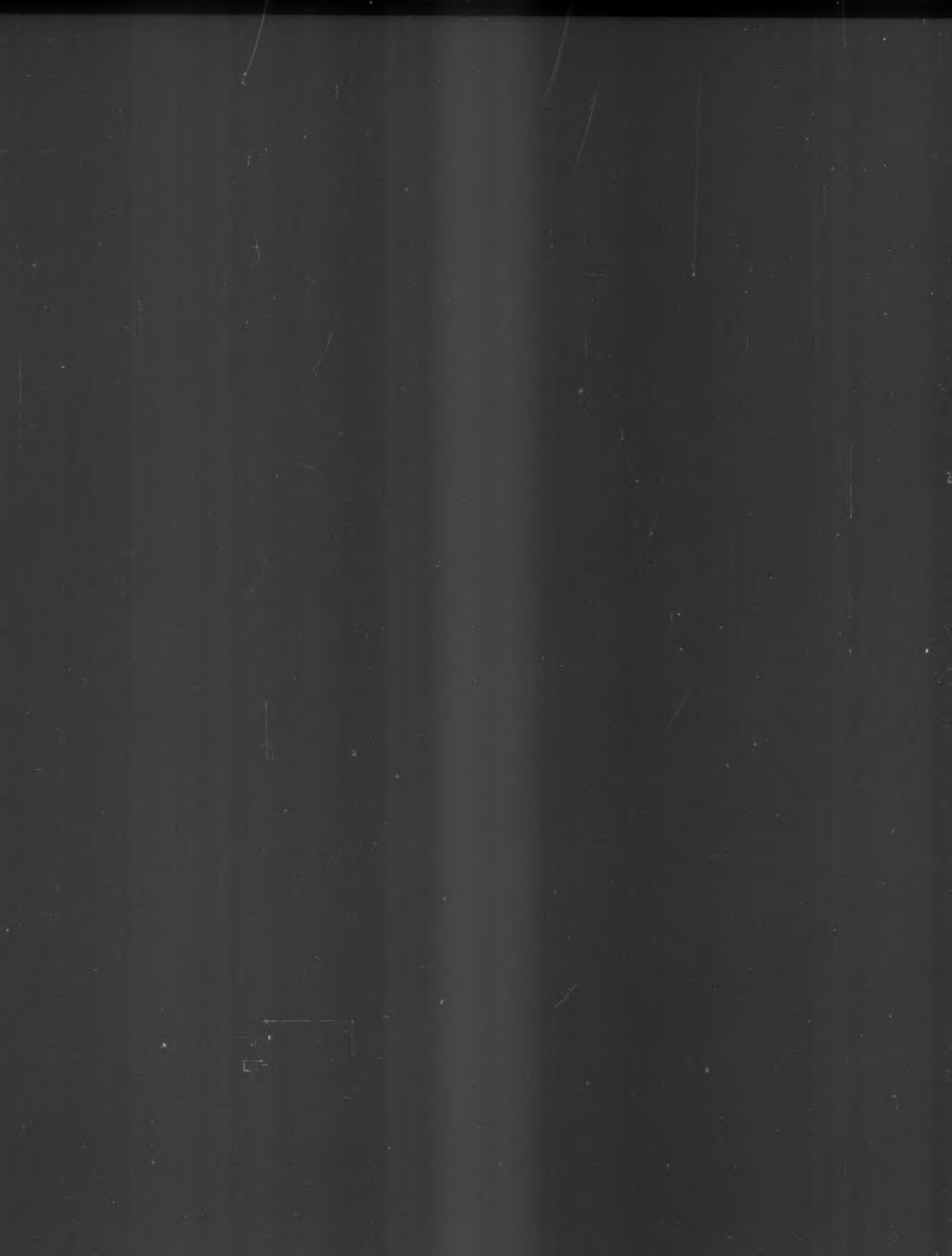
City _____ State_____ Zip_____

Charge my credit card ☐ VISA ☐ MasterCard ☐ Deposit Account

Credit Card No._____

Expiration Date _____

Signature _____

A Publication of the National Cancer Institute

NIH Publication No. 85-22